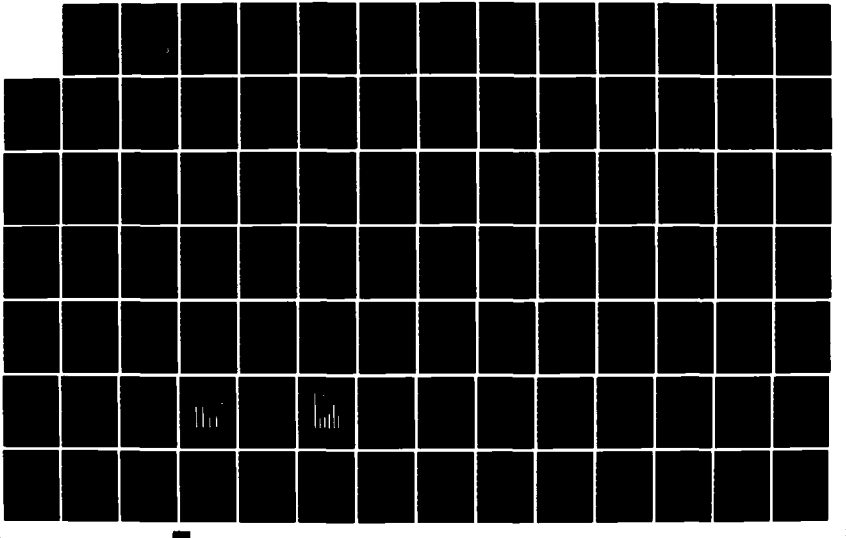
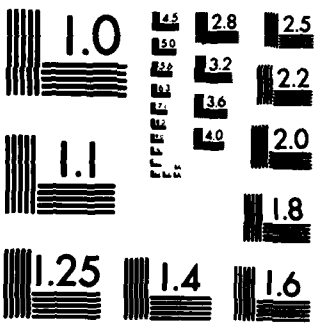
AD-A146 838 SUBJECTIVE AND OBJECTIVE EVALUATION OF PITCH EXTRACTORS 1/2
FOR LPC AND HARMO..(U) BOLT BERANEK AND NEWMAN INC
CAMBRIDGE MA V R VISWANATHAN ET AL. JUL 84 BBN-5726
UNCLASSIFIED DCA100-83-C-0023 F/G 20/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0

Bolt Beranek and Newman Inc.

(35)

Report No. 5726



# Subjective and Objective Evaluation of Pitch Extractors for LPC and Harmonic Deviations Vocoders

## Final Report

July 1984



Prepared for:
Defense Communications Agency

BBN Report No. 5726

# SUBJECTIVE AND OBJECTIVE EVALUATION OF PITCH EXTRACTORS FOR LPC AND HARMONIC DEVIATIONS VOCODERS

Final Report
Contract No. DCA100-83-C-0023

Authors: V.R. Viswanathan and W.H. Russell

July 1984

Prepared by:

Bolt Beranek and Newman Inc.
10 Moulton Street
Cambridge, MA 02238

Prepared for:

Defense Communications Agency
Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, VA 22090

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>BBN Report No. 5726 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>SUBJECTIVE AND OBJECTIVE EVALUATION OF PITCH EXTRACTORS FOR LPC AND HARMONIC DEVIATIONS VOCODERS | | 5. TYPE OF REPORT & PERIOD COVERED<br>Final Report<br>Feb. 1983 - July 1984<br>6. PERFORMING ORG. REPORT NUMBER<br>BBN Report No. 5726 |
| 7. AUTHOR(s)<br>Vishu R. Viswanathan and William H. Russell | | 8. CONTRACT OR GRANT NUMBER(s)<br>DCA100-83-C-0023 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Bolt Beranek and Newman Inc.<br>10 Moulton Street<br>Cambridge, MA 02238 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Defense Communications Agency<br>Contract Management Division, Code 680<br>Washington, D.C. 20305 | | 12. REPORT DATE<br>July 1984<br>13. NUMBER OF PAGES<br>121 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Pitch extraction, pitch detector, voiced/unvoiced detection, pitch and voicing errors, subjective evluation of pitch extractors, objective evaluation of pitch extractors.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes the work performed on subjective and objective evaluation of pitch extractors for use in 2.4 kbit/s LPC and harmonic deviations(HDV) speech coders. From a review of the available pitch extractors, five algorithms were chosen and implemented. An existing algorithm was modified to extract automatically accurate, reference pitch and voicing data from the subglottal signal recorded simultaneously with the speech

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

signal using a miniature accelerometer; this data was used for objective evaluation of pitch extractors. Since the accelerometer is relatively insensitive to acoustic background noise, this method yields accurate pitch and voicing data even in noise.

For formal subjective evaluation of the chosen pitch extractors, a speech database of 48 sentences that are likely to cause pitch and voicing errors was developed. Test stimuli were generated using two 2.4 kbit/s coders (LPC and HDV), 6 pitch extractors (5 algorithms under test and the reference), and 2 noise conditions (clear and Air-Borne Command Post or ABCP noise). Two separate tests, one for each noise condition, were run. Eight listeners rated the speech quality of the stimuli on an 8-point scale. The results of the subjective tests showed the reference subglottal-signal-based pitch extractor to be the best under all four coder/noise conditions, validating its use as reference in the subsequent objective evaluation work. The results also indicated two best pitch extractors under test; one produced the highest mean rating in the clear and the other, in ABCP noise.

The objective evaluation method developed in this work involves comparing, on a frame-by-frame basis, the test pitch extractor data with the reference pitch data, computing objective pitch and voicing error measures, and averaging over the sentences from the speech database. For developing objective measures, a study was first conducted to assess the perceptual effects of introducing different types and amounts of pitch and voicing errors into the reference pitch data. Based on the results of this study, a large number of objective measures for evaluating pitch extractors were developed, using different combinations of one or more of the following components: percentage of the frames containing voicing errors and gross pitch errors, energy weighting, weighting based on the duration of the errors, pitch frequency and pitch error weighting, and context-dependent error measurement. Two previously reported objective measures were also implemented. Twelve of the objective measures developed in this work provided consistently high correlation with mean subjective ratings in each of the four cases, two coders each in clear and in ABCP noise. In contrast, the previously reported measures provided high correlation in the clear and substantially lower correlation in the noise. Finally, the best overall objective meaure produced excellent correlations, ranging from -0.953 to -0.995, with the overall mean subjective rating. This measure also predicted nearly perfectly the rank ordering of the five test pitch extractors by the subjective rating, in all coder/noise conditions.

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## ACKNOWLEDGMENTS

The authors wish to thank their colleagues K. Field for bringing up some of the pitch extraction programs on the VAX computer and for analyzing the results of the subjective tests; A.W.F. Huggins for his help in the design of the subjective tests and in the analysis of the results of the subjective tests; and A. Derr for his help in the development of the speech database. The following individuals provided their pitch extraction programs for evaluation in this project: B. Dupree and N. Green, Joint Speech Research Unit, U.K.; C. Gillman, University of Wisconsin; W. Henke, Belmont, MA; E. Singer, Lincoln Laboratory; and L.F. Willems, Institute for Perception Research, The Netherlands. G. Doddington and B. Secrest of Texas Instruments provided their speech and hand-edited databases, and J. Picone also of Texas Instruments provided the details of the TI's objective pitch evaluation measure. Finally, the authors would like to thank J. Lambert and G. Moran of the Defense Communications Agency for their interest and encouragement during the course of this project.

## A SPECIAL NOTE

We had received magnetic tape copies of the pitch extraction programs directly from the respective authors or their associates. We carefully tested these programs before we evaluated them using subjective and objective methods. In cases of problems, we consulted with the authors whenever possible. However, we do not rule out the possibility of mistakes in the way we had used these pitch extraction programs. We have pointed out in the report an inadvertent error in the amount of delay we had used for the Gold pitch detector, which led to its higher pitch and voicing errors and lower subjective scores. We sincerely apologize for this mistake. A thorough check did not reveal any further mistakes. We have confirmed that the same sets of pitch and voicing data were used for both subjective and objective evaluation. This ensures that the high correlation scores we obtained for our objective pitch evaluation measures are indeed accurate.

# 1. INTRODUCTION

The overall objective of this project was to conduct a comparative evaluation of selected pitch and voicing extraction algorithms for use in 2.4 kbit/s LPC and harmonic deviations speech coders. In this chapter, we state the specific goals of this work (Section 1.1), present the highlights of this work (Section 1.2), and provide an overview of the rest of the report (Section 1.3).

## 1.1 Goals of the Project

As part of an earlier project (Contract No. DCA100-80-C-0039), we had developed the harmonic deviations (HDV) coder [1, 2]. At a synchronous transmission data rate of 2.4 kbits/s, the HDV coder produces noticeably better speech quality than does the U.S. Government standard coder LPC-10. However, both the HDV and LPC-10 coders use the same AMDF-DYPTRACK algorithm for extracting the pitch and voicing data [3]. In our experience dealing with the real-time LPC-10 coder implemented on the MAP-300 array processor [4] and in the experience of others, the AMDF-DYPTRACK algorithm produces pitch and voicing errors for certain types of speakers. Also, for a given speaker, the algorithm works well when the speaker talks with a nearly monotone pitch, but tends to produce pitch errors when the speaker uses a large pitch range to

reflect his/her excitement, for example. The pitch and voicing errors mentioned above cause the coder output speech to degrade noticeably. For the HDV coder, pitch errors produce an additional effect. Since the HDV coder achieves speech quality improvement over the LPC-10 by correcting the amplitudes of the LPC model spectrum at a selected set of the harmonics of the fundamental frequency, any error in the extracted pitch tends to reduce the effectiveness of the spectral corrections and hence reduce the extent of improvement over the LPC-10.

Based on the above considerations, our goal in the present project was to study and comparatively evaluate existing pitch and voicing extraction algorithms, with specific emphasis on their use in the HDV coder. For comparison purposes, we also included in our work the application to the LPC coder. Specific objectives of this project are stated as follows:

- Study and review a number of published algorithms for pitch and voicing extraction
- Select several algorithms for comparative evaluation and implement them on our computer
- Develop a speech database containing speech materials and speakers specifically chosen for testing pitch extraction algorithms
- Develop a method for obtaining accurate pitch and voicing data to be used as reference in objective evaluation of pitch extractors

- Evaluate the selected pitch and voicing extraction algorithms using formal subjective listening tests in which subjects compare the speech outputs obtained using each of these algorithms in both the HDV and LPC coders, under two conditions:

  1. clean or noise-free input speech and

  2. input speech corrupted by Air-Borne Command Post (ABCP) noise

- Develop objective measures for evaluating pitch extractors, which produce high correlation with subjective judgments from the above-mentioned listening tests.

Before we present the highlights of our work, we point out that we use the term 'pitch extractor' to mean pitch and voicing extractor. Also, strictly speaking, pitch frequency refers to a perceived attribute of the physically measurable quantity fundamental frequency. For the purpose of this report, however, we do not distinguish between the two terms, pitch frequency and fundamental frequency.

## 1.2 Highlights of the Work

From a review of the available pitch extractors, we chose and implemented five algorithms: one is the AMDF-DYPTRACK pitch algorithm used in LPC-10 (denoted in this report as AMDFD); one is a time-domain, parallel-processing algorithm (Gold pitch detector); one uses the power spectrum to determine the

harmonic pattern (Harmonic-Sieve or H-S algorithm); and two use the cepstrum (Interactive Laboratory System or ILS algorithm and Joint Speech Research Unit or JSRU algorithm). To bring up the five algorithms on our computer, we obtained the magnetic tape copy of the working programs from the respective authors or their associates. In our initial tests of the five algorithms, we used Texas Instruments speech and hand-edited pitch databases. Also, we conducted informal listening tests on speech output from the LPC and HDV coders using each of the five pitch extractors.

To extract accurate, reference pitch and voicing data, which is required for objective evaluation of pitch extractors, we developed and tested an automatic method that uses as input the subglottal signal recorded simultaneously with the speech signal using a miniature accelerometer. The method (denoted in this report as FPRDM, where FPRD stands for "fundamental period" and M stands for "modified") is a modification of the one originally developed at Massachusetts Institute of Technology. The original algorithm provides pitch-synchronously a pitch value and an associated confidence level. Our modified algorithm extracts the voicing decision and provides both pitch and voicing data time-synchronously as required by the LPC and HDV coders and as required for objective evaluation. Since the accelerometer is relatively insensitive to acoustic background noise, the FPRDM method yields accurate

pitch and voicing data even in noise.

For formal subjective evaluation of the chosen pitch extractors, we developed a speech database of a total of 48 sentences from three male and three female speakers, representing a wide range of pitch. We selected the speech materials from a phoneme-specific database of 120 sentences (developed as part of an earlier BBN project) and the speakers from a population of 12 males and 12 females in such a way that both the sentences and the speakers are likely to cause pitch and voicing errors, which facilitates efficient subjective and objective testing of the pitch extractors. We used the real-time LPC-10 coder running on the MAP-300 in this selection process. We generated the test stimuli using two coders (LPC and HDV), six pitch extractors (FPRDM and the five algorithms under test), and two noise conditions (clear and ABCP). We ran two separate tests, one for each noise condition. Eight listeners rated the speech quality of the stimuli on an 8-point rating scale. We computed the mean rating scores for each pitch extractor under four conditions: LPC/Clear, HDV/Clear, LPC/Noise, and HDV/Noise. The major results of the subjective tests are as follows:

- The FPRDM pitch was judged to the best for all four coder/noise conditions. This result validates our use of the FPRDM pitch as reference for objective evaluation.

- Of the five test pitch extractors, the AMDFD algorithm produced the best speech quality in the clear condition, with the JSRU method being slightly worse. In ABCP noise, however, the JSRU method was far superior to all four pitch extractors; AMDFD was rated third, being only slightly worse than the second place ILS algorithm. Overall, the results show JSRU and AMDFD to be the two best pitch extractors.

- Differences in the mean ratings between the HDV coder and the LPC coder were small. The reason for this result is that the sentences included in the tests were designed to expose the differences among the pitch extractors and are not suited to demonstrate the speech quality differences between the two coders.

For developing objective pitch evaluation measures, we first conducted a study involving informal listening tests, to assess the perceptual effects of introducing different types and amounts of pitch and voicing errors into the reference pitch data. The objective evaluation method involves comparing, on a frame-by-frame basis, the test pitch data obtained using a pitch extractor under evaluation with the reference FPRDM pitch data, computing objective pitch and voicing error measures, and averaging over the sentences from the speech database. Based on the results of the above-mentioned perceptual study, we developed a large number of objective measures for evaluating pitch extractors, using different combinations of one or more of the following components: percentage of the processed frames containing voicing errors and pitch errors that are larger than a threshold (10%), weighting of the errors based on speech signal energy, weighting based on the duration of consecutive

errors, weighting based on pitch frequency and pitch error, and context-dependent error measurement. We also implemented two objective measures previously reported in the literature. We correlated the data from each objective measure with the mean subjective rating scores in each of eight different conditions: For each of the four cases, LPC/Clear, HDV/Clear, LPC/Noise, and HDV/Noise, we considered the subjective rating data in two ways, once as the overall mean ratings over the complete database of 48 stimulus sentences (six speakers x eight sentences) and once as the more detailed mean ratings over eight subsets of six stimulus sentences each. The correlation values evaluated at the detailed level will be lower than those evaluated at the overall level. From the correlation results, we selected a set of 12 objective measures each of which produced consistently high correlation in all eight conditions. The mean correlation over the four coder/noise conditions ranged from -0.906 to -0.982 at the overall rating level and from -0.842 to -0.902 at the detailed level. The mean correlation over all eight conditions ranged from -0.891 to -0.942. In contrast, the two previously reported measures produced high correlation in the clear and substantially lower correlation in the ABCP noise. The mean correlations over the eight conditions were only -0.561 and -0.824 for those two measures. Finally, our best overall objective measure produced correlations ranging from -0.953 to -0.995 at the overall level and from -0.867 to -0.929 at the

detailed level. This measure also predicted nearly perfectly the rank ordering of the five test pitch extractors by the subjective rating, in all four coder/noise conditions.

## 1.3 Overview of the Report

In Chapter 2, we describe briefly the 2.4 kbit/s LPC and HDV coders used in this work. Chapter 3 contains a description of the five pitch extractors we chose to evaluate. In Chapter 4, we present the results of our initial objective and informal subjective evaluation of the chosen five pitch extractors, using Texas Instruments speech and hand-edited pitch databases. In Chapter 5, we present a method for generating accurate, reference pitch and voicing data; this method uses subglottal signal recorded during speech with a miniature accelerometer attached to the speaker's throat. Chapter 6 deals with formal subjective evaluation of pitch extractors, and it contains a description of the speech database we designed, the subjective test we used, and the test results we obtained. In Chapter 7, we present the results of our effort to understand the perceptual effects of pitch and voicing errors, as a precursor to the development of objective measures for evaluating pitch extractors. The topic of objective evaluation of pitch extractors is then treated in Chapter 8. Finally, in Chapter 9, we present a summary of this

work and discuss some issues that warrant further research.

## 2. TWO 2.4 KBIT/S SPEECH CODERS

Below, we describe briefly the 2.4 kbit/s LPC and HDV coders we used in this project. Both coder simulations permit the option to read in pitch and voicing data from disk files generated using separate pitch programs. This option allowed us to generate coder output speech data for each of various pitch extractors in a convenient manner.

### 2.1 LPC Coder

The analog input speech is lowpass filtered at 5 kHz, sampled at 10 kHz, and divided into non-overlapping frames of 20 ms duration for linear prediction and pitch analyses. For every analysis frame, the following quantities are transmitted using a total of 48 bits: a synchronization bit, voicing status (1 bit), pitch (6 bits, logarithmic quantization), speech signal energy (5 bits, logarithmic quantization), and 12 log area ratios (35 bits total, optimal scalar quantization involving orthogonal transformation of the log area ratio vector and nonuniform scalar quantization of transformed parameters [1, 2]). The receiver uses the binary pulse/noise excitation for the all-pole synthesis filter to generate the output speech.

## 2.2 HDV Coder

A detailed description of the 2.4 kbit/s HDV coder algorithm is given in [1, 2]. The HDV coder algorithm may be summarized as follows. In the transmitter, the analog speech is lowpass filtered at 5 kHz, sampled at 10 kHz, and divided first into non-overlapping frames of 20 ms duration and then into 9-frame blocks. A variable frame rate (VFR) algorithm is used to select and transmit only 6 frames of data every block, along with a block header (6 bits long) to identify the transmitted frames. For every frame selected by the VFR algorithm, the following quantities are transmitted: a synchronization bit, voicing status (1 bit), pitch (6 bits, logarithmic quantization), speech signal energy (5 bits, logarithmic quantization), 12 log area ratios (37 bits total, optimal scalar quantization), and 3 selected spectral deviations (2 bits each) between the log spectrum of the speech signal in the frame and the log spectrum of the all-pole model. These quantities are quantized, coded, partially error-protected, and transmitted across the channel. Error protection is provided by sending the block header in 3 copies, sending the frame voicing bit in 5 copies, and using 3 Hamming (7, 4) codewords per transmitted frame to protect 12 selected bits of the parameter data. At the receiver, the data for the untransmitted frames are regenerated by linear interpolation between adjacent transmitted frames. The

output speech of the coder is synthesized, pitch-synchronously for voiced frames and every 10 ms for unvoiced frames, by generating the excitation signal using the spectral deviations and applying it to the all-pole synthesis filter.

The AMDF-DYPTRACK pitch algorithm used in the HDV coder algorithm produces pitch estimates that are already quantized to one of 60 levels. The HDV coder program, therefore, did not quantize the pitch. For investigating the use of other pitch extractors in the HDV coder, we modified the original HDV coder program by adding provisions to quantize pitch using 6 bits. We used a pitch coding range of 20 to 156 samples or 64 to 500 Hz, which is about the same range used in the AMDF-DYPTRACK method. We quantized the log pitch using a method developed at BBN, which makes effective use of the quantization levels at the small pitch period end [5].

## 3. FIVE PITCH EXTRACTORS

Extraction of pitch and voicing forms an important part of a variety of speech processing systems. As testimony to this fact, literally hundreds of algorithms have been reported in the literature for extracting pitch and voicing. For a detailed survey of these algorithms, the reader is referred to the book [6]. Most algorithms for extracting pitch and voicing have three components: preprocessor, basic extractor, and postprocessor. The basic extractor performs the actual measurement of pitch and voicing. The function of the preprocessor is to process the input speech signal with the goal of simplifying the foregoing measurement task. The distribution of the algorithm complexity between the preprocessor and the basic extractor varies over individual algorithms. The postprocessor smooths or detects and corrects possible errors in the extracted pitch contour.

From a review of the existing pitch extractors, we chose for our comparative evaluation work five algorithms that are described briefly in this chapter. Our choice was governed to some extent by our attempt to include as many different pitch and voicing extraction approaches as possible and to a large extent by the ease of transportability of the algorithm implementation to our VAX-11/780 computer (running under VMS operating system). Of the

chosen five algorithms, one is a time-domain algorithm (Gold); one uses the so-called average magnitude difference function (AMDF-DYPTRACT); one uses the power spectrum (Harmonic-Sieve); and two use the cepstrum (ILS and JSRU algorithms). The AMDF-DYPTRACK and ILS algorithms were already available on our computer. To bring up the other three algorithms on our computer, we obtained the magnetic tape copy of the working programs from the respective authors or their associates and made only minor changes as described below. In our installation, each pitch program accepts as input a speech waveform file and provides as output a frame-by-frame pitch data file containing zero for unvoiced frames and pitch period in number of samples for voiced frames.

### 3.1 AMDF-DYPTRACK Algorithm

As noted above in Section 1.1, the AMDF-DYPTRACK algorithm is used in the U.S. Government standard coder LPC-10 [3] and in the HDV coder we developed as part of an earlier project [1, 2]. This algorithm computes an estimate of the pitch for a frame by locating the minimum of the so-called average magnitude difference function (AMDF) and uses a dynamic-programming-based tracking (DYPTRACK) method to refine and smooth the computed pitch estimates. We shall refer to this algorithm by the abbreviation AMDFD.

As preprocessing, the AMDFD algorithm lowpass filters the input speech at 800 Hz and spectrally flattens the filtered signal by passing it through a second-order linear prediction inverse filter. An estimate of the pitch period is determined by computing the AMDF function for the inverse-filtered signal over a set of sixty lags (over the range of 20 to 156 samples) and identifying the lag at which the AMDF function is minimum. Notice that this pitch estimation process involves an inherent quantization as it allows only sixty lags. The voicing detector uses an energy measure, a zero-crossing count, and the maximum to minimum ratio of the AMDF function. As postprocessing, the pitch and voicing results are smoothed and corrected by a dynamic programming algorithm, which introduces two frames of delay.

We believe that the Fortran AMDFD program we have on our computer corresponds to Version 44 of the LPC-10 coder. The AMDFD algorithm as implemented in LPC-10 has been "hard-wired" to operate under the conditions of LPC-10 such as 8 kHz sampling rate and 22.5 ms frame rate. Also, the algorithm extracts one pitch value and two half-frame voicing decisions each frame. Our simulations of both the HDV and LPC coders accept only one voicing decision per frame. Furthermore, some of the decision parameters used in the algorithm are adaptive in that their values evolve continuously in time. Thus, the algorithm will not, in general, produce satisfactory results if one

uses it on a one-sentence-at-a-time basis rather than for continuous speech processing. To resolve these problems and to be able to use the algorithm for different frame rates and two sampling rates (8 and 10 kHz), we made several modifications to the algorithm as part of an earlier project. These modifications are described in detail in the report [1].

### 3.2 Gold Pitch Detector

In this method [7], a series of measurements are made on the peaks and valleys of a lowpass-filtered speech signal to produce six separate functions. Each of these six functions is processed by a simple pitch estimator. The resulting six pitch period estimates are analyzed using a decision logic to determine the pitch period. The decision logic is set up to give one pitch period estimate per frame period. The degree of agreement among the six simple pitch detectors is a parameter that is used in making the voicing decision. We note that the parallel processing method of Gold and Rabiner [8], which has been used in other studies comparing pitch and voicing extraction algorithms [9, 10], is a simplified version of the Gold pitch detector.

The version of the Gold pitch detector used in our project was

implemented in C by Mr. E. Singer of Lincoln Laboratory. This version included a 3-point median smoothing of the extracted pitch data. Initially, this program required 8 kHz sampling rate and 22.5 ms frame rate. Upon our request, Mr. Singer provided us with a modified version to use either 8 kHz or 10 kHz sampling rate and any user-specified frame rate.

### 3.3 Harmonic-Sieve Method

In this method [11], input speech is lowpass filtered to a bandwidth of 2.5 kHz. Power spectrum of the filtered signal is computed over a 40-ms analysis frame. Peaks are located on the power spectrum as potential harmonics of the fundamental frequency. Using a harmonic-sieve procedure, the algorithm determines the harmonic pattern and the associated pitch frequency that best match the measured spectral peaks. The extent of the match is used in making the voicing decision. The harmonic-sieve (H-S) method uses no postprocessing. The authors of the H-S method argue that their method is optimal in that it is based on Goldstein's theory of pitch perception in complex sounds [12].

We received a magnetic tape copy of a Fortran implementation of the H-S method from Mr. L.F. Willems of the Institute for Perception Research,

Eindhoven, The Netherlands. This program assumes 10 kHz sampling rate and 100 frame/s analysis rate. We modified the program to use an analysis rate of either 100 or 50 frames/s.

### 3.4 ILS Cepstral Algorithm

This algorithm is part of the Signal Technology Inc. Interactive Laboratory System software package [13]. Input speech is preemphasized, and the log magnitude spectrum is computed. A tapered cosine window is applied to the log magnitude spectrum before computing the cepstrum. The cepstrum is weighted using a cepstral multiplier, and the peak of the weighted cepstrum is located. The cepstral lag corresponding to the peak gives the estimated pitch period. Voicing is extracted using a statistical linear discriminant function approach involving the following quantities: the cepstral peak value, the number of zero crossings in the frame, the first reflection coefficient resulting from linear prediction analysis of input speech over the frame, and the linear prediction residual signal energy. A heuristic method is used to smooth the extracted pitch over 3 frames; this method produces a delay of 1 frame. In our work, we used the default parameter settings given in the ILS package.

### 3.5 JSRU Cepstral Algorithm

This algorithm was developed at the Joint Speech Research Unit, Gloucestershire, U.K. [14]. In this algorithm, input speech is sampled at 10 kHz, preemphasized using simple differencing (6 dB/octave preemphasis), and analyzed over 512-sample frames at a rate of 100 frames/s. Analysis includes a number of steps: Hamming windowing; power spectrum computation; computing a ratio of low-frequency power (40-1200 Hz band) to high-frequency power (2.7-3.9 kHz band); log power spectrum computation; "conditioning" the log power spectrum by replacing the values at the high-frequency end (above about 4.1 kHz) with a constant equal to the average over the baseband (20 Hz to 4.1 kHz) and by eliminating excessive dips in the baseband below this average; cepstrum computation from the conditioned log power spectrum; smoothing the cepstrum using a 3-point FIR filter with weights 1, 2, and 1; locating the peaks in the cepstrum; and finding the location and values of the two largest cepstral peaks. The next step involves the use of a set of heuristics in making the voicing decision and in obtaining a pitch value for the frame under consideration. The heuristics examine the size of the cepstral peaks and the low-frequency to high-frequency power ratio mentioned above; check for possible pitch frequency doubling or halving; and ensure consistency with the pitch and voicing data over the past two frames. Finally, the algorithm

declares any isolated unvoiced frame as voiced.

For the 100 frame/s analysis rate, the algorithm introduces five frames of delay: three frames of delay caused by an offset of the input speech frame from the center of the 512-sample analysis interval, one frame of delay introduced in the heuristics module, and another frame of delay caused by an isolated unvoiced frame check. Two additional frames of delay are introduced for the voicing state to allow for smoothing in the JSRU synthesizer. We removed the latter two-frame delay of the voicing state and compensated for the five frames of delay (by shifting) prior to output to a pitch file. We also modified the JSRU program to allow the use of either 100 frame/s or 50 frame/s analysis rate. For the latter analysis rate, the delay introduced by the algorithm is only three frames.

# 4. INITIAL INVESTIGATION OF THE FIVE PITCH EXTRACTORS

The purposes of our initial investigation of the five pitch extractors reported in the last chapter were to ensure the proper operation of the individual pitch extractors, devoting attention particularly to the changes we made to the original algorithms; to examine carefully the delay introduced by each algorithm; to get some initial reading on the comparative performance of the five algorithms; and to conduct some preliminary testing of the use of these pitch and voicing algorithms in the LPC and HDV coders. For this investigation, we used a subset of the speech database developed by Texas Instruments (TI), as described in Section 4.1. We also used several simple objective error measures given in Section 4.2. The results of our objective and subjective tests are presented, respectively, in Sections 4.3 and 4.4.

## 4.1 TI Speech and Pitch Databases

We obtained from TI a speech database of a total of 58 sentences from 32 male and 26 female speakers (one sentence per speaker) ranging in age from 6 to 87 years [10]. Speech was digitized at 12.5 kHz. We also received from TI reference pitch files (10 ms frame) for these 58 sentences, which were obtained by hand-editing the pitch data generated using the ILS cepstral pitch extractor.

For use in our investigation, we selected a subset of 12 sentences from speakers ranging in age from 7 to 80 years, and digitally resampled the 12.5 kHz waveform files at 10 kHz using the interpolation-decimation approach. To check the accuracy of the TI hand-edited pitch we examined waveform displays of the 12 sentences. We compared the voicing status and pitch period values as determined by visual inspection with the TI pitch data. We found that the TI data gave good estimates of pitch period values in steady state voiced regions and represented the pitch dynamics reasonably well when pitch changed rapidly. We also noted that the TI voicing decisions were correct for all obviously voiced and unvoiced regions. At transitions, the TI data had a tendency to extend voicing somewhat. Although a few of the voicing decisions could be questioned, we concluded that the TI hand-edited pitch data provided reasonably accurate pitch period estimates and voicing decisions.

For the objective and subjective evaluations reported in Sections 4.3 and 4.4, we used only 6 of the chosen 12 sentences. Table 1 gives the five distinct sentences (one sentence spoken by two speakers), and Table 2 gives speaker details.

For the full 58-sentence database, we also obtained from TI the pitch files generated using their integrated correlation pitch program [15]. We

1. Very few angels are always wise and pure.
2. A great future is always provided the student of music.
3. Almost everything involved making the child mind.
4. The view of the present will largely be reached in the following century.
5. The wife's figure had already adjusted by itself.

TABLE 1. Speech materials used in the chosen subset of the TI database

| Sex | Age | Spoken Sentence # |
|---|---|---|
| Male | 24 | 1 |
| Male | 36 | 5 |
| Male | 42 | 3 |
| Female | 33 | 4 |
| Female | 36 | 5 |
| Female | 40 | 2 |

TABLE 2. Details of speakers included in the chosen subset of the TI database.

could not get their pitch program for proprietary reasons. This pitch algorithm, which we refer to as the TI pitch algorithm, uses an adaptive 1-pole filter for preprocessing the input speech, a modified correlation technique for extracting candidate pitch values, and a dynamic programming technique for both making voicing decision and obtaining a smoothed pitch

estimate [15]. We included the TI pitch algorithm in some of our investigations, primarily for comparison purposes.

### 4.2 Pitch and Voicing Error Measures

In objective evaluation of a pitch extractor under test, we compare the test pitch data obtained using this extractor with the reference pitch data on a frame-by-frame basis. A comparison of the test pitch value with the reference pitch value for any given frame indicates one of four possibilities listed below.

1. Both the test and the reference pitch values are zero indicating that the frame was declared unvoiced in both test and reference pitch files. For this case, no error has occurred.
2. The reference pitch value is non-zero, but the test pitch value is zero. Thus, a voicing error has occurred, and we denote this error as a voiced-to-unvoiced (VUV) error.
3. The reference pitch value is zero, but the test pitch value is non-zero. Thus, a voicing error has occurred, and we denote this error as an unvoiced-to-voiced (UVV) error.
4. Both the reference and test pitch values are non-zero. For this case, we compute the pitch error between the two values.

For the fourth case, we classify the pitch error as a gross pitch error if the magnitude of the quantity 100(FT-FR)/FR exceeds a prespecified

threshold, where FT and FR are, respectively, the test and reference pitch frequency in Hz. In our investigation, we have used a threshold of 10% in deciding gross pitch errors. A pitch error that is not a gross pitch error is called a fine pitch error.

We developed an interactive program, called PEVAL (short for pitch evaluation), to compare test pitch data with reference pitch data and compute pitch and voicing error statistics. PEVAL uses our command interpreter software so that the user may interactively control the execution of various components of the program. The PARAMETER command allows the user to set various parameter values, and the COMPARE command allows the user to compare the test pitch data with the reference pitch data for one utterance or a group of utterances, by providing as input one or more pairs (test, reference) of pitch files. (For additional PEVAL commands, see Chapter 8.) We note that the frame sizes of the reference and test pitch files need not be equal. In fact, PEVAL performs the comparison at any user-specified frame size by converting, if necessary, the reference and test pitch data to correspond to this frame size via linear interpolation of log pitch.

PEVAL computes the count of each of the three types of errors, VUV, UVV, and gross pitch errors, as a percentage of the total number of frames used in

the comparison, and determines the total error by adding the three percentages. For fine and gross pitch errors, the program computes the mean and standard deviation. Also, for gross pitch errors, the program identifies pitch frequency doubling and halving errors and computes the total number of each. If the magnitude of the difference between the reference pitch frequency and half (twice) the test pitch frequency, expressed as a percentage of the reference pitch frequency, is less than a threshold (we used 10%), the pitch error is classified as pitch frequency halving (doubling). PEVAL computes various other statistics including duration of a consecutive occurrence of a given error type, location of error (voiced region, unvoiced region, voiced-unvoiced transition, and unvoiced-voiced transition), and missing voiced or unvoiced regions. (See Chapter 8 for more details.)

The five basic error measures (percent VUV error, percent UVV error, percent gross pitch error, and mean and standard deviation of fine pitch error) have been previously used for objective evaluation of pitch extractors [9]. However, we point out that pitch error is computed in [9] as difference in pitch period in number of samples between test and reference cases and is compared against a threshold in deciding if it is a gross pitch error. We believe that using the percentage pitch frequency error as described above is perceptually more relevant. Also, this method allows us to

compare directly two pitch files generated using different speech sampling rates.

### 4.3 Objective Evaluation

We processed the 12 TI sentences mentioned in Section 4.1 through each of the five pitch extractors, using 10 kHz sampling rate and 10 ms frame size (i.e., a frame rate of 100 frames/s), and evaluated the resulting pitch data using PEVAL with TI hand-edited pitch data as reference. We found that the pitch and voicing errors given by PEVAL were considerably higher for the Harmonic-Sieve and Gold pitch extractors than for the other three. This result prompted us to check if we were using the correct time delay for each pitch extractor. To do this task, we modified PEVAL to include the option of skewing the test pitch file with respect to the reference pitch file by a prespecified number of frames. If there was an unaccounted delay being introduced by the pitch extractor, the total error (sum of VUV, UVV, and gross pitch errors) should decrease as this delay is removed. The results of this test on the various pitch extractors are shown in Fig. 1. (A negative skew refers to removing frames from the beginning of the test file, thus shifting the test file backward with respect to the reference file.)

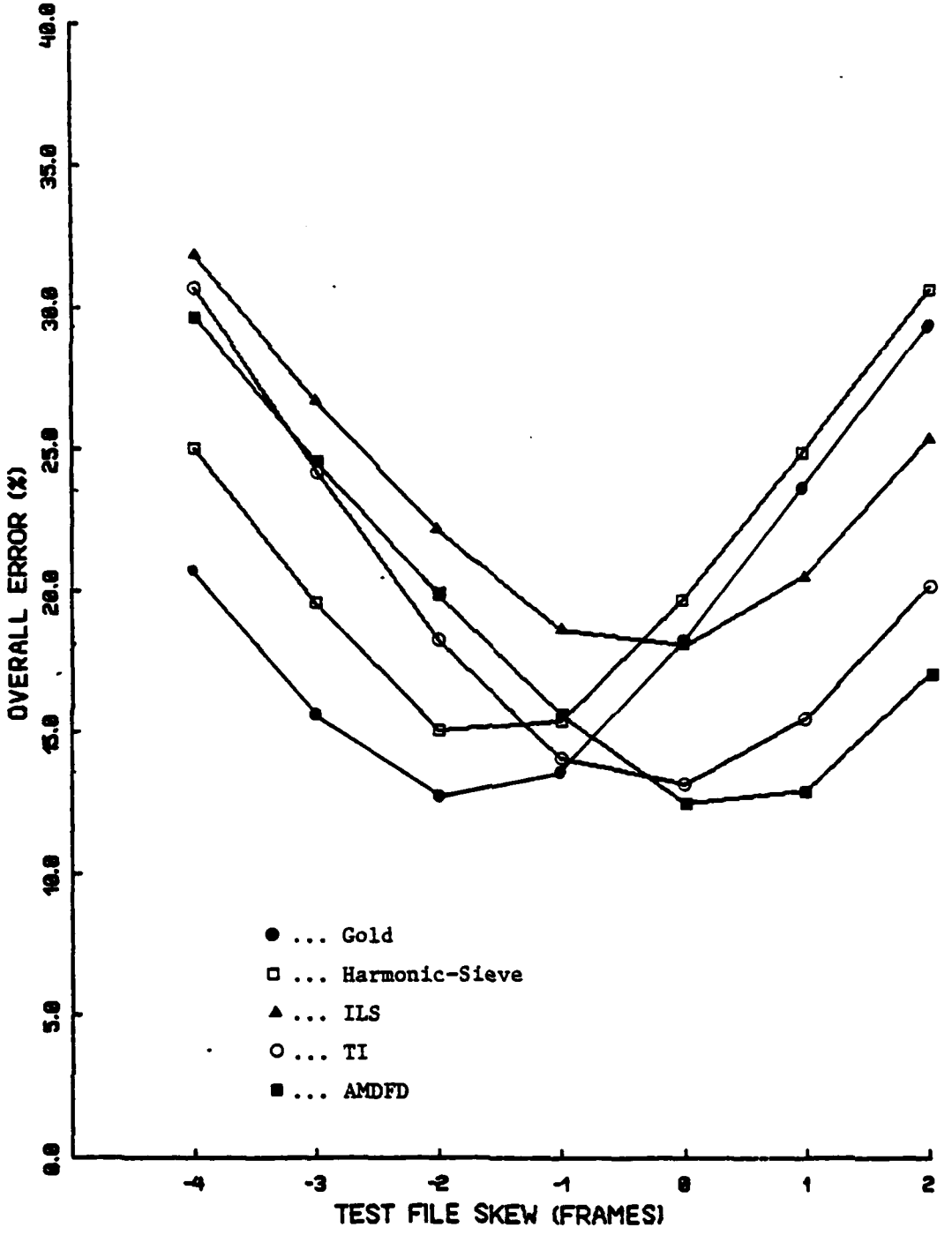



FIG. 1. Total error plotted as a function of the number of frames of skew between the test and the reference pitch files, for each of five pitch extractors.

We see from Fig. 1 that the plot for the Harmonic-Sieve method has a minimum between -1 frame and -2 frames. We examined the Fortran code of this method and found that the 40-ms analysis interval used contained the current 10-ms frame and the three past frames. If we associate the extracted pitch with the center of the analysis interval and the current frame's pitch with its center, we find a delay of 15 ms or 1.5 frames. As PEVAL uses only integer number of frames, we decided to use a 20 ms or two-frame delay. We note that the documentation of the Harmonic-Sieve method did not mention about any delay.

Referring to Fig. 1, we find that the Gold algorithm shows a minimum at -2 frames. From a discussion with the author of the program (E. Singer), we identified the two frames of delay as being caused by lowpass filtering and median smoothing.

Notice from Fig. 1 that all other pitch extractors yield a minimum at a skew of 0 frames, which indicates that we had correctly accounted for their delays. (The plot for the JSRU method, not shown in Fig. 1, also yielded a minimum at 0 frames of skew.) The compensation of the foregoing delay reduced the total error by about 5% for the Harmonic-Sieve method and by about 10% for the Gold algorithm.

With the time delay correctly compensated as discussed above, we evaluated, using PEVAL over the subset of six TI sentences given in Section 4.1, all six pitch extractors: AMDFD, Gold, H-S, ILS, JSRU, and TI. The resulting various error measures are given in Table 3. Except for fine error mean, all other errors are each expressed as a percentage over the total number of data frames considered (see Section 4.2). Total error is again the sum of VUV, UVV, and gross errors.

| Error | AMDFD | Gold | H-S | ILS | JSRU | TI |
|---|---|---|---|---|---|---|
| Percent VUV Error | 2.81 | 7.60 | 9.44 | 3.46 | 7.26 | 0.79 |
| Percent UVV Error | 1.45 | 2.29 | 0.99 | 3.36 | 2.10 | 4.67 |
| Percent Gross Error | 6.77 | 2.93 | 5.17 | 5.82 | 1.90 | 6.36 |
| Total Error | 11.03 | 12.82 | 15.60 | 12.64 | 11.26 | 11.82 |
| Pitch Doubling | 0.40 | 0.00 | 2.43 | 0.10 | 0.00 | 0.05 |
| Pitch Halving | 0.35 | 1.54 | 0.05 | 2.30 | 1.10 | 2.68 |
| Fine Error Mean | 0.52 | 1.13 | -0.01 | -0.33 | 0.59 | 0.83 |

TABLE 3. Pitch and voicing error results obtained over the six TI sentences, for six pitch extractors.

From Table 3, we see that over the six sentences considered, the AMDFD algorithm produced the least error and the Harmonic-Sieve method produced the most error. The JSRU algorithm, which was only slightly worse than AMDFD, yielded the least gross pitch error. AMDFD provided the least voicing error (sum of VUV and UVV errors).

To test the effect of smoothing, we applied a 3-point median smoother to the pitch data from the various pitch extractors and reexamined their errors using PEVAL. The median smoother was designed to work continuously on all frames regardless of voicing boundaries and thus was able to correct one frame isolated voicing errors. The smoothing did not decrease the overall error for any of the algorithms by more than 0.5%, and in the cases of the Harmonic-Sieve and Gold algorithms, the error was actually increased by approximately 0.1%. A possible reason for the increase in the Gold algorithm is that it already uses 3-point median smoothing as noted above.

### 4.4 Subjective Evaluation

We performed informal listening tests of the six TI sentences of speech synthesized using the HDV coder with the pitch and voicing data from each of the six pitch extractors. Judging from the overall speech quality, we felt

that the AMDFD algorithm was the best, closely followed by the JSRU, TI, ILS, and Gold algorithms. The Harmonic-Sieve method produced the worst quality; specifically, both voicing and pitch doubling errors were quite audible. The ILS algorithm produced all of the pitch halving errors in two sentences, which were quite evident in informal listening. For the JSRU method, the HDV coder speech sounded quite natural during correctly voiced regions, but the voicing errors significantly degraded the overall quality. In general, the presence of UVV errors degraded the speech quality less than did the presence of VUV errors. The reason for this result is that the frame energy associated with UVV errors is in general lower than that of the VUV errors.

# 5. A METHOD FOR GENERATING REFERENCE PITCH DATA

In this chapter, we review Henke's FPRD (short for fundamental period) algorithm for extracting accurate pitch from the subglottal accelerometer signal recorded during speech (Section 5.1). We then describe a conversion routine we developed to extract voicing decision from FPRD output data and to convert pitch-synchronous FPRD pitch data to time-synchronous data as required by our LPC and HDV coders and as required for objective evaluation (Section 5.2). The results obtained using this modified FPRD program on the speech signal are presented in Section 5.3.

## 5.1 FPRD Algorithm

This algorithm was developed at MIT by Dr. W. Henke [16]. In this method, a two-channel tape recording is made of the speech signal transduced by a microphone and the subglottal signal transduced by a miniature, lightweight accelerometer, which is attached with double-sided adhesive tape to the speaker's throat on the midline in the suprasternal notch and just below the glottis. We used the Vibro-Meter Corporation (formerly BBN Instruments Corporation) Model 501 accelerometer, which weighs less than 2 grams. The FPRD method uses the accelerometer signal to extract accurate pitch data. Figure 2 displays the speech signal and the accelerometer signal,

during phonation of the vowel [a]. From the figure, it is clear that the subglottal signal displays the individual pitch periods (albeit shifted by about 1 ms relative to the speech signal, because of propagation time delay), without the resonances of the vocal tract. It has been found that the time of the major negative-going zero crossing in the subglottal signal, shown by arrows in Fig. 2, provides a stable segmentation point for delimiting individual pitch periods [16]. Henke refers to the rapid change around this zero crossing from outward to inward acceleration immediately following the maximum acceleration as the "flyback stroke". The flyback stroke occurs at or shortly after the instant of glottal closure.

Given the accelerometer signal, the FPRD method locates the zero crossing associated with the flyback stroke by identifying signal maxima and minima and using heuristics, and provides pitch-synchronously a pitch value and a voicing confidence level. The latter quantity takes the integer values 1 to 4, with 1 indicating least confidence and 4 indicating most confidence.

We make several observations. First, the FPRD program was developed originally for making pitch period and jitter measurements [16]. Second, this method is being used in a computer-based system of speech-training aids for the deaf [17, 18]. Experience gathered in this application has suggested that

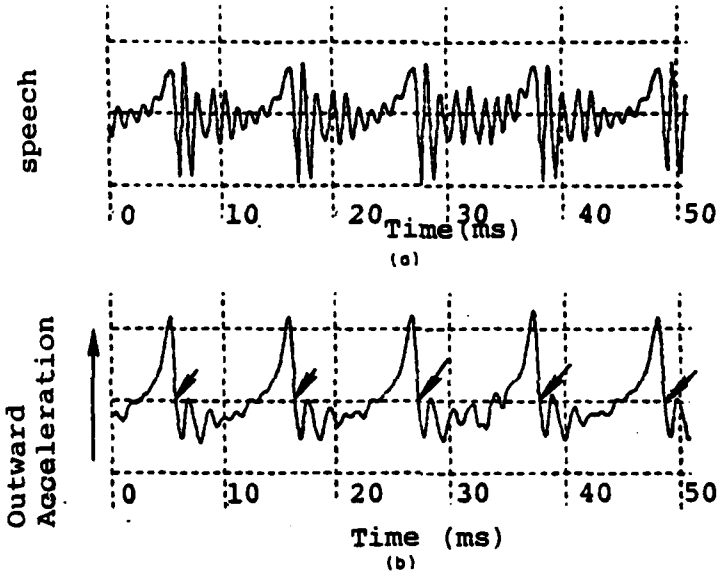



Fig. 2 (a) Signal from audio microphone 10 cm from lips, vowel [a].
(b) Simultaneous signal from an external accelerometer attached to the throat just below the glottis.
(Figure taken from [16].)

the FPRD algorithm operates reliably over a range of subjects including adult males, adult females, and children. Third, notice that the FPRD algorithm does not provide the binary (voiced/unvoiced) voicing decision required by most narrowband vocoders. Fourth, the accelerometer is essentially insensitive to acoustic background noise at low frequencies and only mildly sensitive at high frequencies. This property implies that the FPRD algorithm can be used to extract accurate pitch even in acoustic background noise. In fact, as part of another BBN project, the accelerometer has been used in conjunction with a noise-cancelling microphone to transduce noise-immune speech signal [19, 20]. In the same project, it has been found that of the various accelerometer positions on the head and neck, the position just below the glottis provides the highest spectral amplitude at frequencies around the pitch frequency, which makes this position best for pitch extraction [19, 20].

We obtained a listing of the source FPRD program from Henke. The program was in a structured high-level processor language that was not available to us. Fortunately, we received a Fortran version of the FPRD program from C. Gillman at the University of Wisconsin. We brought up this program on our VAX/VMS computer by making the required Fortran syntax changes and by incorporating our file input/output software. The FPRD program requires the user to specify two input parameters: the speaker's average pitch frequency

in Hz and the accelerometer signal polarity (see Section 5.2). We initially tested and debugged the program using the speech signal as input, even though the program was designed for the subglottal signal input. We processed the same speech file through our program and through the original FPRD program at MIT. The outputs from the two runs were found to be identical.

## 5.2 Voicing Decision and Time-Synchronous Pitch

We interpreted the voicing confidence level output from the FPRD program as follows: A value of 1 indicates a "definitely unvoiced" period; a value of 4 indicates a "definitely voiced" period; and values of 2 and 3 indicate transition periods. We mention that for a confidence level of 1, the FPRD program provides as pitch estimate the average pitch period. (Recall that the average pitch frequency is one of the user-specified inputs.) To check the validity of our interpretation given above and to develop a technique of assigning a binary voicing status to the transition periods, we processed through the FPRD program five sentences of the accelerometer signal from a male speaker. The accelerometer signal and the speech signal were previously recorded simultaneously on a two-channel tape recorder and digitized using our two-channel A/D facility as part of another contract effort at BBN [19, 20]. A visual display of the accelerometer signal and the speech signal was

examined to determine the location of the glottal events specified by the FPRD program output and to identify the correspondence between the accelerometer signal and speech signal events. An example of the display is shown in Fig. 3. The waveform at the top of the figure is a section of the accelerometer signal, and the corresponding section of the speech waveform is shown at the bottom of the figure. The FPRD program positions the epoch boundary at the zero-crossing of the "flyback stroke". The arithmetic sign of the slope of the "flyback stroke" in the accelerometer signal determines the signal polarity parameter referred to earlier in Section 5.1.

The results of the foregoing investigation confirmed the validity of our interpretation of confidence levels 1 and 4 as, respectively, unvoiced and voiced. For level 4 cases, the pitch estimates from FPRD were found to be quite accurate. Also, we identified a simple method of assigning a binary voicing status to the transition periods: Declare all transition periods that occur in the middle of an unvoiced region (a region with consecutive confidence levels of 1) as unvoiced and declare all other transition periods as voiced. The voiced transition periods can thus occur immediately preceding, succeeding, or in the middle of a voiced region (with consecutive confidence levels of 4). We hasten to point out that this simple rule worked well over the five sentences we investigated, but caused some voicing errors

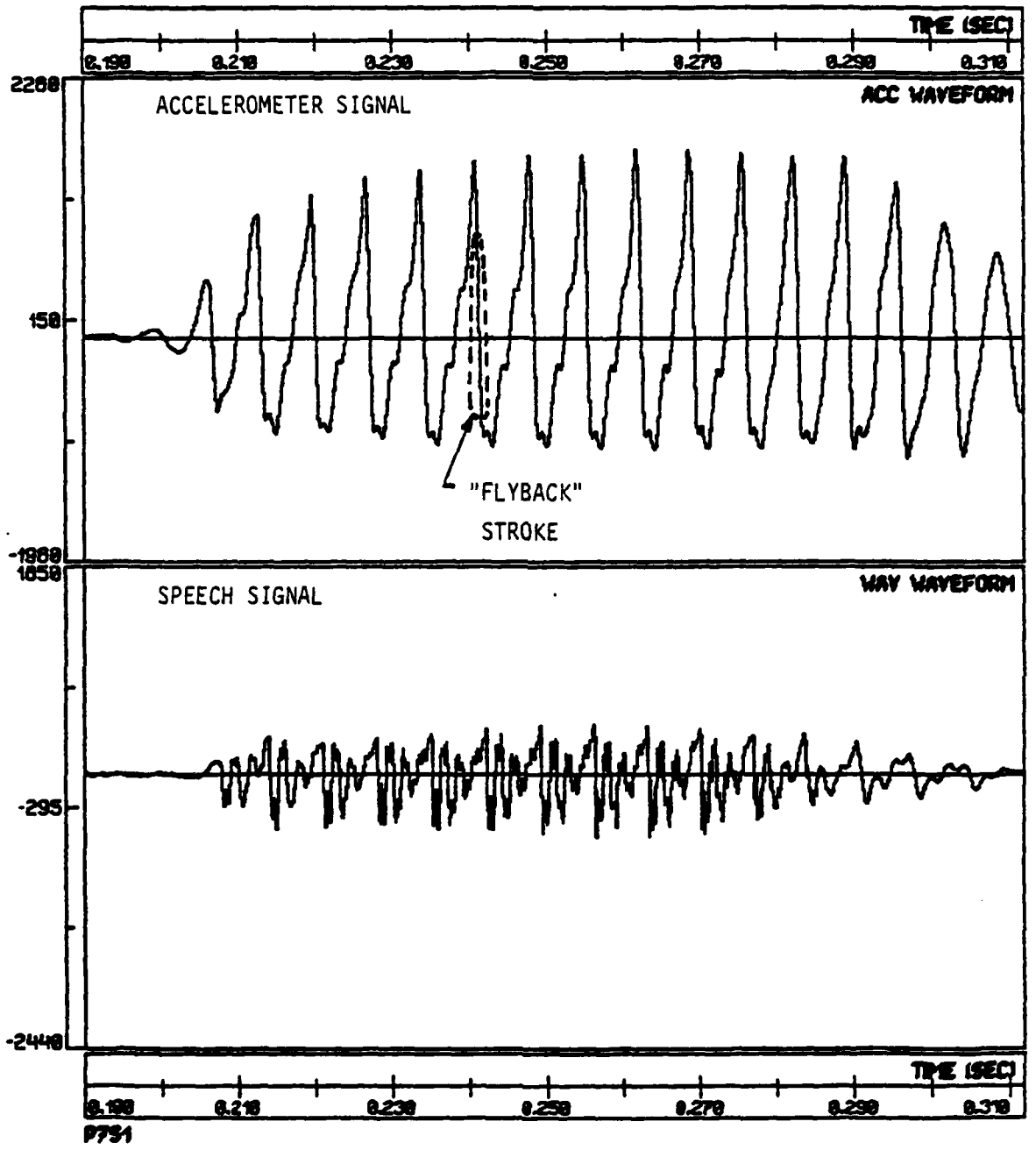



Fig. 3 Waveforms of the accelerometer and speech signals.

over a larger speech database used in our formal subjective tests. Refinements to the voicing decision rule are given in Section 6.2.

We incorporated into the PEVAL program a subroutine to determine the binary voicing decision for the FPRD pitch and to convert the pitch-synchronous pitch data to time-synchronous (or frame-by-frame) data at a frame rate specified by the user. This subroutine determines the pitch value for a frame as the pitch-synchronous pitch period that occurs at the center of the frame. Notice that it does not perform any interpolation. We refer to the resulting time-synchronous FPRD pitch data with binary voicing decision as FPRDM (M stands for "modified") pitch data. It is the FPRDM pitch that we used in all our subsequent subjective and objective evaluation work.

We processed the foregoing five sentences of speech through the HDV coder twice, once using the AMDFD pitch and once using the FPRDM pitch. Through informal listening tests, we found that the FPRDM pitch produced more natural-sounding speech than did the AMDFD pitch.

## 5.3 Performance with Speech Signal as Input

To test how well the FPRDM program extracts pitch with speech signal as input, we processed the six TI sentences (see Section 4.1) through the FPRDM program with 10-ms frame size and compared the resulting pitch data, using PEVAL, with the reference TI hand-edited pitch data. The error results are given in Table 4.

| Error | FPRDM |
|---|---|
| Percent VUV Error | 15.53 |
| Percent UVV Error | 0.05 |
| Percent Gross Error | 3.42 |
| Total Error | 19.00 |
| Pitch Doubling | 0.00 |
| Pitch Halving | 0.10 |
| Fine Error Mean | 0.95 |

TABLE 4. Pitch and voicing error results obtained over the six TI sentences, for FPRDM.

Comparing the results given in Table 4 with those given in Table 3 for six other pitch extractors, we find that the VUV error was unacceptably large

for FPRDM. All other error measures were in fact quite low for FPRDM. To examine the VUV errors caused by FPRDM in more detail, we observed visual displays of several speech waveforms and noted the locations of voicing transitions. The observed transition locations were compared with the pitch-synchronous voicing confidence level outputs of the FPRD pitch extractor. These comparisons showed that the FPRD pitch extractor yielded a confidence level of 1 or a definitely unvoiced decision for several moderate and low energy regions of voiced speech. We synthesized several sentences with the HDV coder using the FPRDM speech-derived pitch as input. Informal listening tests showed that the speech for FPRDM pitch was more natural in voiced regions than speech for AMDFD pitch. However, many raspy and hoarse-sounding effects were present in the speech for FPRDM pitch confirming the presence of obvious voicing errors. The overall speech quality produced by the FPRDM pitch was better than that produced by the Harmonic-Sieve pitch, even though the total error was larger for FPRDM than for Harmonic-Sieve (19.0 vs 15.6).

We emphasize that the FPRD program was not designed to use the speech signal as input. We believe that the performance of the FPRDM program on the speech signal can be improved substantially using a better (perhaps a separate) voicing detector.

# 6. FORMAL SUBJECTIVE EVALUATION OF PITCH EXTRACTORS

Below, we present in Section 6.1 our design of the speech database for use in subjective and objective evaluation of pitch extractors. In Section 6.2, we describe the generation of the reference pitch data using FPRDM and the test pitch data using the five pitch extractors reviewed in Chapter 3. The design of the subjective tests is treated in Section 6.3, and the results from the subjective tests are presented in Section 6.4.

## 6.1 Speech Database

The acceptability of a pitch extractor for use in the LPC or HDV coder depends on the frequency with which it generates pitch or voicing errors that degrade the perceived quality of the coder output speech. For reliable and efficient testing of pitch extractors, the speech database used must therefore contain a substantial number of speech events that are likely to generate pitch and voicing errors. Test utterances that fail to create pitch and voicing errors do not provide useful data for classifying and comparing candidate pitch extractors, and should therefore be excluded from the database. In developing a speech database, we must also use properly chosen speakers. The pitch range of the speakers is an important factor [21]. Also, it has been our experience and the experience of others that some voices

present severe difficulties to pitch extractors.

To determine the speech utterances and speakers that are likely to generate pitch and voicing errors, we processed speech through the real-time LPC-10 coder running on the MAP-300 array processor [4] and evaluated the presence of pitch and voicing errors by listening to the coder output. As speech material, we used a subset of a phoneme-specific database of about 120 sentences developed as part of an earlier BBN project [5]. We used a number of speakers and three listeners in this investigation. Sentences containing unvoiced consonants caused most pitch and voicing errors: presence of unvoiced stops was particularly effective in causing errors. Sentences with only voiced consonants caused the fewest errors. Using the results of this test, we chose a set of 51 sentences. We selected 12 male speakers and 12 female speakers with a wide range of pitch. We recorded the accelerometer signal and the speech signal using a two-channel tape recorder as each of the 24 speakers read the 51 sentences.

To begin selection of the final database, we first informally listened to all of the sentences spoken by all of the speakers processed through the real-time LPC-10 coder. On the first pass, we selected, for each speaker, the sentences that caused noticeable pitch or voicing errors. On the second pass,

we evaluated the severity of the errors for only the sentences selected in the first pass. From this information we were then able to select the six speakers who produced the most (and most severe) errors. An additional criterion we used in selecting the speakers was to span a wide range of pitch, from low-pitched males to high-pitched females. The six speakers we selected are 3 females (LW, BF, and MA) and 3 males (AW, DG, and PH).

The final step in the database selection process was to find the sentences for these speakers that caused the most pitch and voicing errors. To enable us to make direct comparisons among the speakers for specific utterances, we selected sentences that caused errors for all six chosen speakers. We also selected sentences each containing only specific types of speech sounds, in the hope that these sentences would cause different types of pitch errors to occur. In addition, we selected, for each speaker, two to four other sentences that caused errors specifically for that speaker. The final database we selected contains a total of 50 sentences. The first six sentences given in Table 5 were recorded from all six speakers and the remaining sentences were recorded only from the speakers identified within parentheses.

After selecting the database, we digitized the speech and accelerometer

| Sentence | Type of Sounds |
|---|---|
| 1. Why were you away a year, Roy? | Voiced |
| 2. Patty cut up a potato cake. | Unvoiced stops |
| 3. Which tea party did Baker go to? | Stops, affricates |
| 4. Chip took a picture. | Unvoiced stops, affricates |
| 5. Whose shaver has three fuses? | Fricatives |
| 6. A thickset officer pitched out her hash. | Unvoiced |
| 7. Take a copy to Pete. (AW) | Unvoiced stops |
| 8. Pat talked to Kitty. (AW,LW) | Unvoiced stops |
| 9. Keep quiet at church. (BF,MA,PH) | Unvoiced stops, affricates |
| 10. Katie typed a paper. (BF,MA) | Unvoiced stops |
| 11. Peter took out a potato. (DG,LW) | Unvoiced stops |
| 12. Teacher taped up a packet. (DG) | Unvoiced stops, affricates |
| 13. Teacher patched it up. (DG) | Unvoiced stops, affricates |
| 14. Quite quiet at church. (PH) | Unvoiced stops, affricates |
| 15. A thief saw a fish. (PH) | Fricatives |

TABLE 5. Sentences used in the speech database.

signals for the selected sentences using a two-channel digitizing technique, which preserved the time-alignment of the two signals. The resulting waveform files were then edited and split into two files: one containing the speech signal and the other containing the corresponding accelerometer signal.

We then proceeded to add digitally ABCP noise to clean speech sentences to generate the noisy speech database. From the sponsor-supplied tape containing sentences of speech recorded in an ABCP noise environment, we digitized the noise-only parts and digitally "spliced" these together to

obtain one long noise file with about 7.5 seconds of noise. We listened to this file to verify that it did not contain any obvious repetitious patterns or pops and clicks because of the splicing process. The noise file sounded about the same as we heard on the sponsor's tape.

We added the ABCP noise to individual speech sentences to obtain a prespecified signal-to-noise ratio (SNR), as follows. We computed the average per-sample energy of the noise. For speech, we computed per-sample energy in 10-ms frames over the given sentence, identified the frames with energies above the 90th percentile, averaged the peak energy over these frames, and subtracted a constant (we used 5 dB) to obtain a robust estimate of the average speech signal energy. This method is robust as it is not as sensitive to the presence of pauses and silence in speech as is the overall average energy. From the per-sample average energies of the noise and the speech files, we scaled the noise so as to produce a specified SNR over each sentence.

We added the noise to several utterances from our speech database with various SNR's and performed informal listening tests to decide which SNR best matched the SNR of the sentences on the sponsor's tape. We found that an SNR of 7 dB gave the best match. Upon adding noise to several test utterances we

noted that the adjustment of the noise level for each utterance, to maintain the SNR of 7 dB, caused small changes in background noise from sentence to sentence that could be perceived in informal listening tests. In an actual ABCP noise environment, the noise level would not change as speaking levels changed. We therefore decided to add a fixed noise level to all sentences. This fixed level was obtained by averaging the noise levels required to produce a 7 dB SNR for a number of sentences. We thus generated a 50-sentence ABCP noise-added speech database.

## 6.2 Generation of Reference and Test Pitch Data

We generated pitch files for the two sets of 50 sentences of speech corresponding to the clear and noisy databases, for each of the five pitch extractors: AMDFD, Gold, Harmonic-Sieve, ILS, and JSRU. We used a frame rate of 50 frames/s required by the 2.4 kbit/s LPC and HDV coders. We then generated the pitch-synchronous pitch data using the FPRD program on the accelerometer signal files for the 50 sentences of speech, and converted the data, using PEVAL, to time-synchronous pitch and voicing data (FPRDM), initially at 10-ms frame size. We treated the resulting FPRDM pitch data as reference in both clear and noisy cases; this is quite reasonable as the accelerometer is essentially insensitive to acoustic background noise. In the

rest of this section, we describe how we carefully examined the FPRDM pitch data and made necessary refinements. For each of the 50 sentences, plots of the speech signal, accelerometer signal, and the frame-by-frame pitch estimates were examined to locate any pitch and voicing errors. High resolution plots were made of voiced regions where the pitch changed rapidly and also of several steady state voiced regions. These plots were used to check the accuracy of the extracted pitch estimates. The confidence level output from the FPRD program and the binary voicing decisions made by our conversion routine were compared with events in the speech and accelerometer signals. We also performed informal listening tests to compare the original speech utterances with synthesized ones that were produced using the FPRDM pitch in the LPC coder. From these tests we concluded that the frame-by-frame FPRDM pitch data was correct for 23 of the 50 sentences in our database. We found at least one instance in each of the remaining 27 sentences where the pitch accuracy or a voicing decision could be questioned.

The utterances that contained errors were reexamined to determine if there were any common characteristics or patterns to the errors that could be detected and corrected by modifications to our conversion routine. Although several types of errors were identified and techniques for correcting them were devised, we were not able to develop, within the scope of this project, a

fully automatic method for obtaining error-free frame-by-frame pitch and voicing decisions from the FPRDM pitch data. Several of the refinements we made to the conversion routine were implemented as options so the user could select the correction techniques that were appropriate for the sentence under examination. Three techniques for making refinements to the voicing decision, discussed below, can be readily included in the automatic conversion routine. The details of the error types and our correction methods are given below.

From our study, we found that accurate frame-by-frame pitch estimates were obtained from the pitch periods classified as "definitely voiced" (confidence level 4). Errors in the frame-by-frame pitch estimates were obtained only when the transition state pitch periods (confidence levels 2 and 3) were used for frame estimates. Recall that our conversion routine considers all transitions state pitch periods as voiced if they occur at the beginning, at the end, or in the middle of voiced regions. The pitch values associated with confidence level 3 were generally reliable; however, the level 2 pitch periods were not. Approximately half of the level 2 transition pitch periods were in error. We modified our conversion routine to substitute the previous level 3 or level 4 voiced pitch period value for each level 2 pitch period that was declared voiced. This scheme worked well for transition state pitch values that occurred in the middle of voiced regions, but was not

appropriate for many of the transition state pitch periods at the end of voiced regions. It was necessary to hand-edit the pitch estimates at the end of several voiced regions.

Most voicing decision errors were caused by inappropriate classification of transition periods by our conversion routine. Regions declared as definitely voiced (confidence level 4) or definitely unvoiced (confidence level 1) were almost always correct. A majority of the voicing decision errors occurred when transition state pitch periods were at the beginning or end of a voiced region. Three modifications were made to the conversion routine to correct a number of the voicing decision errors. First, we noted that if an isolated confidence level of 1 existed at the end of a voiced region followed by a sequence of three or more transition state pitch periods, the transition periods and the isolated level 1 frame must be declared voiced. Second, we also noted that several short voiced regions were not detected by the FPRDM program. Reexamination of the FPRD data showed that each of these regions contained a consecutive sequence of three or more level 3 pitch periods. A provision was added to our conversion routine to declare these regions voiced. Third, we declared any isolated unvoiced frames as voiced with pitch taken from the immediately preceding frame. By application of these three correction methods, we were able to correct approximately 80% of

the voicing decision errors. The remaining voicing errors (about 25 in number) were corrected by hand-editing the pitch files.

All pitch and voicing errors in the 27 utterances were corrected, either by selectively applying the schemes described above or by hand-editing the pitch files. The resulting FPRDM pitch data was used as reference pitch in our subsequent work.

The above discussion might indicate that we made an extensive hand-editing of the FPRDM data. This is simply not true, as will be clear from the facts presented below. First, as we mentioned above, 23 of the 50 sentences did not require any corrections at all. Second, we used the PEVAL program to evaluate the FPRDM pitch data before any corrections were made, with the corrected FPRDM data as reference. For a total of 8,290 10-ms frames analyzed (82.9 seconds of speech), we obtained 93 VUV errors, 37 UVV errors, and 11 gross pitch errors including 1 pitch frequency doubling and 4 pitch frequency halving errors, which represents a total error of only 1.7% errors. In contrast, the five test pitch extractors produced over the six TI sentences total errors in the range of 11 to 16% (see Table 3). (The errors were even larger over our speech database. See Table 6 in Chapter 8.) Third, as we mentioned above, three techniques to correct the errors in the voicing

decision can be readily incorporated into the automatic conversion routine. We implemented these techniques as part of our automatic conversion routine and evaluated the resulting FPRDM data over the 50 sentences. The resulting total error was 1.48%.

The primary motivation for our above-described detailed examination of the FPRDM data was to ensure that it could be used as an accurate reference in our objective evaluation of pitch extractors. We believe that either the original FPRDM data or the one with the additional automatic voicing decision changes would serve well as the intended reference.

### 6.3 Subjective Tests

We decided to conduct formal subjective tests on 2.4 kbit/s LPC and HDV coders using each of the six pitch extractors (AMDFD, FPRDM, Gold, Harmonic-Sieve, ILS, and JSRU), which leads to a total of 12 coding systems. We also considered two acoustic background noise conditions (clear and ABCP noise) for each coding system. From our 50-sentence speech database, we chose for the subjective tests a total of 48 sentences: 8 sentences spoken by each of 6 speakers; six of these sentences are common to all speakers. (The design described below requires that the total number of sentences be an integer

multiple of the number of coding systems.) Use of pairwise comparisons of the 1152 test stimuli (12 coding systems x 2 noise conditions x 48 sentences) would be a formidable task indeed. We therefore decided to adopt a rating test in which a listener rates the overall speech quality of each test sentence on an 8-point scale, with 1 being the worst speech quality and 8 being the best speech quality. It is desirable to limit the duration of each test session to be within 2 hours; otherwise, listeners tend to become tired, loose concentration, and not be consistent in their rating. Guided by this consideration, we decided to run two separate tests, one for clean speech and the other for ABCP noise-added speech. Each test contains 576 (= 12 coding systems x 48 sentences) stimuli, arranged in 12 blocks as explained below.

Since we expected that speech quality differences over the different pitch extractors might often be small, we decided to employ listeners with at least some prior experience in listening to LPC speech. We chose eight people from the BBN Laboratories Speech Group to serve as subjects; two of these eight were closely involved in the preparation of the test stimuli. Since we believe that a dozen or so judgments per stimulus are needed to obtain reasonable average ratings, we decided to run four test sessions for each subject, two versions of the clean speech test and two versions of the noisy speech test. In this manner, we would get, for each test sentence, two

judgments from each of eight subjects or a total of 16 judgments, which should be sufficient. Considering the order in which to run the clean speech tests and the noisy speech tests, we decided to divide the eight subjects into two groups of four subjects each, with the first group going through the four tests, each test run on a different day, in the order Clear I, Noise I, Noise II, and Clear II, and the second group in the order Noise I, Clear I, Clear II, and Noise II. Clear I and Clear II (similarly Noise I and Noise II) involve the same 576 stimuli but use different randomized ordering as discussed below. From the test results, we can evaluate the effect of the ordering of the clean speech and noisy speech tests on the ratings of the listeners. Also, by comparing the Clear I and Clear II as well as the Noise I and Noise II ratings, we can determine how reliable (or consistent) the subjects were in their ratings.

Next, we discuss the method we used for randomizing the order of the test stimuli. The block of 48 sentences (6 speakers x 8 sentences) were first divided into four sub-blocks of 12 sentences each. Each of the 12 sentences in a sub-block was assigned to a particular coding system. We then randomized the ordering of sentences in each sub-block so that no two consecutive sentences were spoken by the same speaker. This randomized ordering made up the first test block of 48 stimuli (4 sub-blocks x 12 sentences). Next, the

coding system ordering was rotated so that the system that was assigned to the first sentence of a sub-block was assigned to the second sentence, the second to the third, and so on. The 12 sentences from each sub-block were then randomized to generate the second block of the test. We repeated this procedure until 12 test blocks were produced, so that all 48 sentences processed by all 12 systems were included. For Clear I and Noise I test tapes, we used different randomization within blocks. We then randomized the ordering of the blocks in the Clear I and Noise I cases to obtain, respectively, the Clear II and Noise II test data.

Finally, for each test, we repeated the first block at the end of the test tape so that each test tape had 13 blocks. The listeners were instructed to use the first block of 48 test sentences in familiarizing themselves with the rating task and with the range of speech quality to be mapped on to the 8-point rating scale. The ratings from this practice block were not used in our analysis.

Using the previously generated pitch files, we generated synthesized speech for our formal listening tests. Each of the chosen 48 test sentences was synthesized using each of the two 2.4 kbit/s coders (HDV and LPC), each of the pitch files (AMDFD, FPRDM, Harmonic-Sieve, Gold, JSRU, and ILS), and each

noise condition (clear and ABCP noise), resulting in 1152 distinct test stimuli. We then prepared test tapes as mentioned above. As we informally listened to the tapes to check if everything was right, we discovered that we had inadvertently substituted, for one of the speakers, the sixth common sentence (see Section 6.1) with a different sentence. In other words, the tapes contained all six common sentences from five speakers and only five from the sixth speaker. Finally, we ran the four tests for each of the two groups of subjects.

Before we present the test results, we must point out that we inadvertently used an incorrect time delay of 60 ms (or three 20-ms frames) for the Gold pitch detector. This led to the Gold pitch extractor's inferior subjective ratings reported below in Section 6.4 and inferior objective scores reported in Chapter 8. (See Subsection 8.1.1 for further discussion).

## 6.4 Test Results

We entered all subjective rating scores into an interactive software facility called RS-1 (a product of BBN Software Products Corporation). The analyses described in this section were performed using the RS-1 system. The plots included in this section were also produced by the RS-1 system.

At the outset, we examined the mean rating score and the standard deviation for each of the eight subjects in each of the four test sessions. The first group of four subjects took the tests (on consecutive days, each test on a different day) in the order Clear I, Noise I, Noise II, and Clear II. As expected, the Clear II mean ratings were in general higher than those for Clear I because the subjects had heard the day before the poorer quality Noise II speech. The second group of four subjects took the tests in the order Noise I, Clear I, Clear II, and Noise II. Since a weekend separated the last two tests, the Noise II mean ratings were not lower than those for Noise I. The mean score and standard deviation varied significantly over subjects and sessions. Therefore, we decided to normalize individual rating scores by subtracting the mean and dividing with the standard deviation computed over the respective subject and session. Next, each subject made two ratings of each stimulus sentence. To assess each subject's reliability, we correlated the two sets of ratings over the 576 stimulus sentences. One subject for the clear condition and three subjects for the noise condition did not produce large enough correlation. We therefore discarded the data from these subjects. The correlation coefficients for the remaining subjects ranged between 0.71 and 0.84. In the rest of our analysis, we thus used 7 subjects (14 ratings per stimulus sentence) for the clear condition and 5 subjects (10 ratings per stimulus sentence) for the noise condition. Also, in computing

the mean scores, we used the average of each subject's two ratings for each stimulus sentence.

Before we present the results comparing the two coders and the six pitch extractors, we mention that subjects were instructed to use the full 8-point rating scale in each test session. Therefore, we caution that the ratings for the noise condition must not be directly compared with the ratings for the clear condition, since subjects were expected to assign different speech quality values, under the two conditions, for a given score. Below, we first present the mean score results over the 48 speaker-sentence combinations and then present the mean score results over the six common sentences for each speaker and over six speakers for each sentence.

Figure 4 shows a bar chart of the mean rating scores over all 48 sentences and over all subjects, comparing the six pitch extractors under each of the four coder conditions: HDV/Clear, LPC/Clear, HDV/Noise, and LPC/Noise. From Fig. 4, we see that the reference pitch FPRDM was judged to be the best for the HDV coder or the LPC coder, under the clear condition or in ABCP noise. In fact, all subjects were in agreement on this point. This result is obviously important for our work on the objective evaluation of pitch extractors, since it validates our use of the FPRDM pitch as reference.

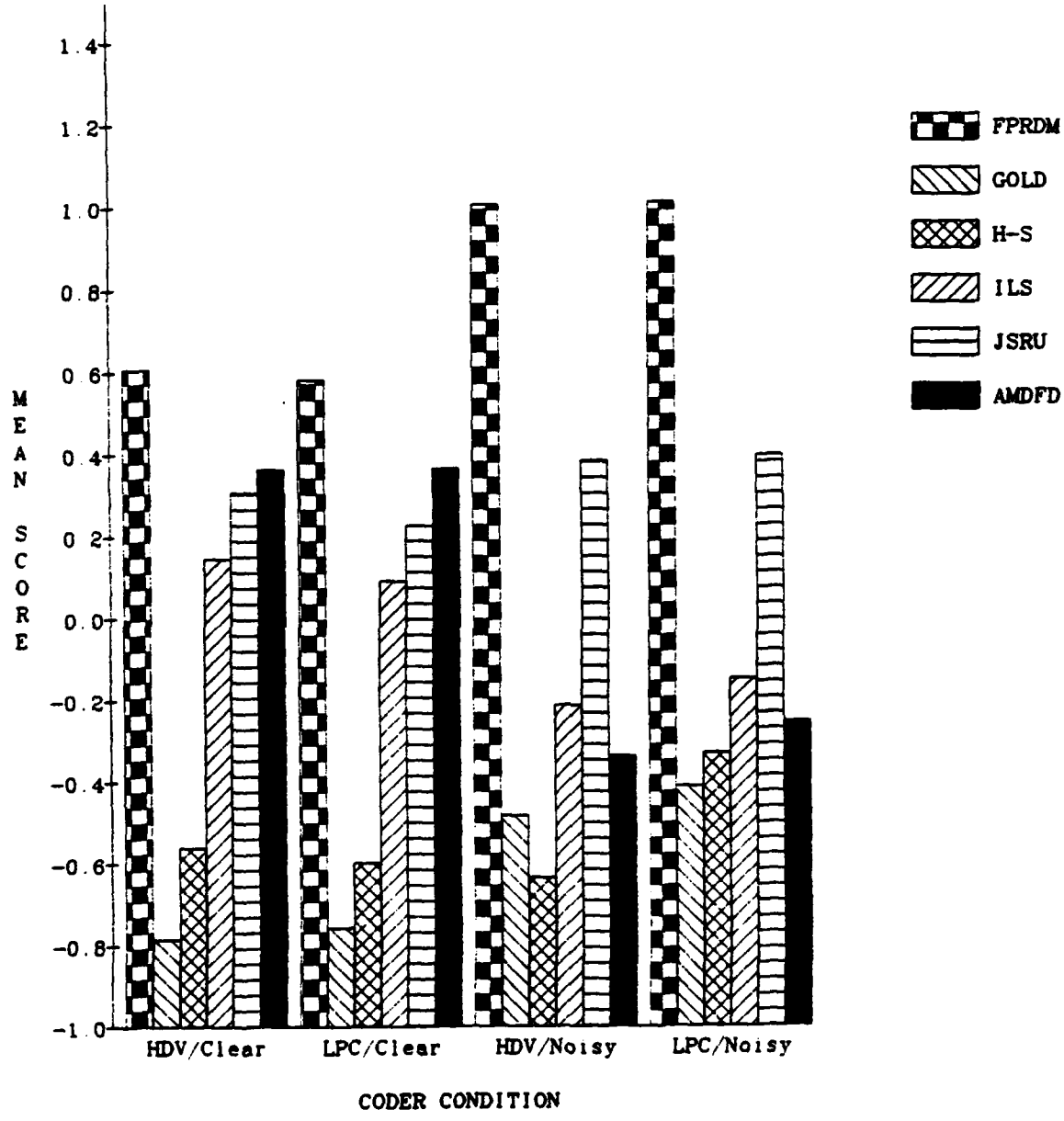



FIG. 4. A bar chart of the mean subjective rating scores, comparing the six pitch extractors under each of the four coder/noise conditions.

Considering the remaining five pitch extractors and the clear condition, we see from Fig. 4 that the AMDFD algorithm produced the best overall speech quality for either of the two coders, with the JSRU method being slightly worse. All seven subjects preferred AMDFD over JSRU for LPC/Clear; 5 subjects preferred AMDFD over JSRU, one had no preference, and one preferred JSRU over AMDFD, for HDV/Clear. The relative ordering of the five pitch extractors for both LPC and HDV coders was, from best to worst, AMDFD, JSRU, ILS, H-S, and Gold. Considering the ABCP noise condition, we observe from Fig. 4 that the JSRU method was far superior to the other four pitch extractors. All five subjects were in agreement on this point. The relative ordering of the five pitch extractors was JSRU, ILS, AMDFD, H-S, and Gold for LPC/Noise and JSRU, ILS, AMDFD, Gold, and H-S for HDV/Noise; AMDFD was only slightly worse than ILS.

Next, we consider the comparison of the HDV coder with the LPC coder. We note that the sentences included in the subjective tests were designed to challenge the pitch extractors and thereby expose the differences among them. These sentences, however, are not particularly suited to demonstrate the speech quality differences between the HDV and LPC coders. The results we obtained were, therefore, mixed in this regard. To make the LPC/HDV comparison a little easier, we have replotted the bar chart in Fig. 5, which

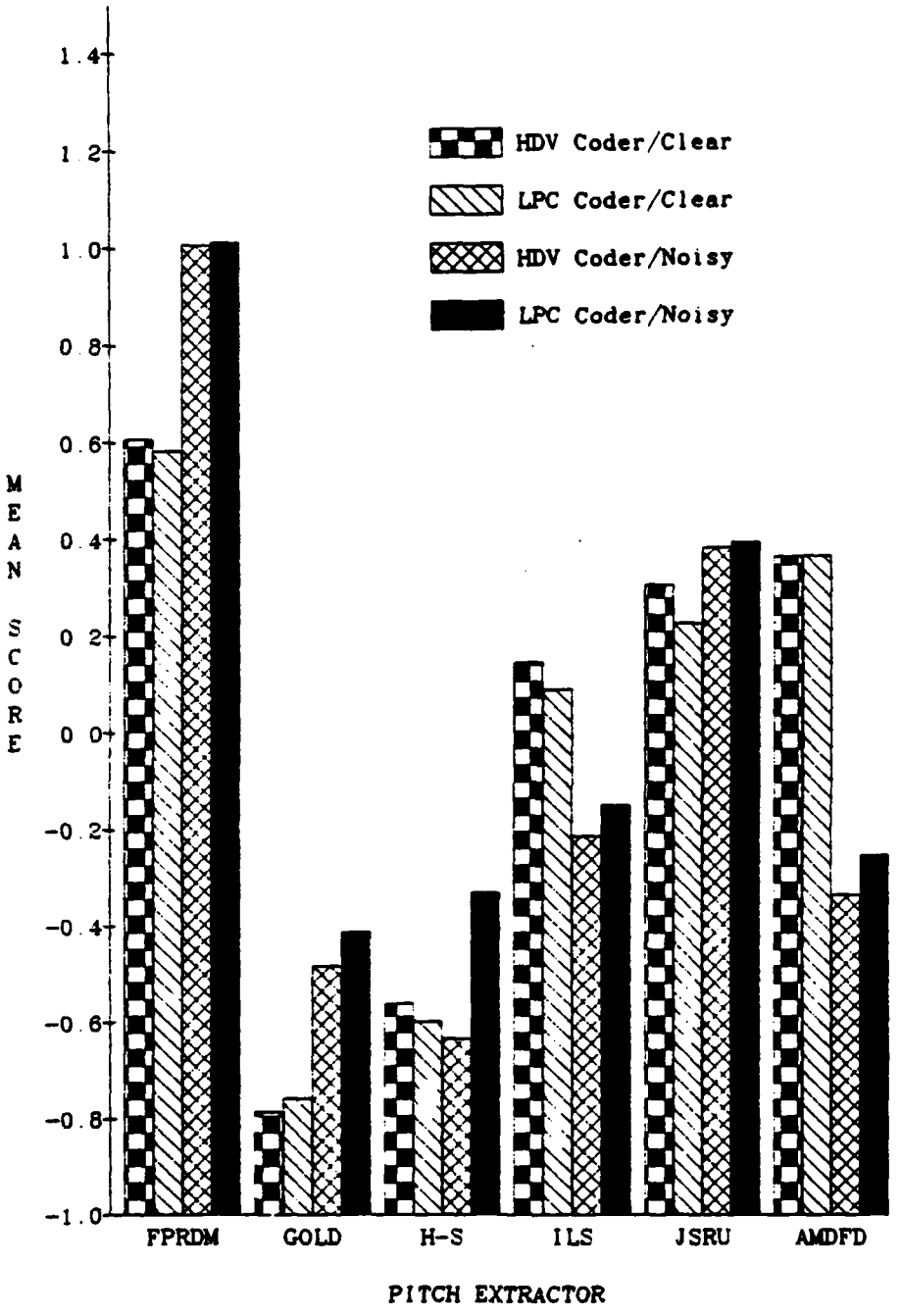



FIG. 5. A bar chart of the mean subjective rating scores, comparing the HDV coder with the LPC coder for each of the six pitch extractors and under clear and noise conditions.

shows the comparison for each pitch extractor. We repeat the caution that we must compare HDV/Clear with LPC/Clear and HDV/Noise with LPC/Noise and not compare between clear and noise conditions. We see from Fig. 5 that the HDV coder was slightly better or about the same as the LPC coder for all but the Gold pitch extractors under the clear condition and that the LPC coder was slightly better or about the same as the HDV coder for all five pitch extractors under the noise condition. (The difference between the two coders in the latter case was, in fact, large for H-S.) For the accurate FPRDM pitch and considering the common six sentences 1-6 (se Table 5), we found that the HDV coder was better over the sentences 1, 5, and 6 and that the LPC coder was better over the sentences 2-4; this result was valid for both clear and noise conditions. Sentences 2-4 contain a number of stops and rapid transitions. We believe that the inferior performance of the HDV coder was in part due to the lower average transmission frame rate employed by the HDV coder (see Section 2.2).

Next, we present the results examining more detailed aspects of the subjective rating data. For this discussion, we have combined the results of the HDV and LPC coders, since the two coders produced similar results as mentioned above and since combining them makes the plots more readable. Also, for computing the mean ratings, we have used, unless said otherwise, only the

data from the 36 speaker-sentence combinations involving all six speakers and the six common sentences. We have included the FPRDM results in plots given below only as a reference, and we make comments on the relative performance for the other five pitch extractors only.

Figures 6 and 7 depict the mean scores for each speaker (we averaged over the six common sentences and all subjects) for the clear and noise conditions, respectively. From Fig. 6, we see that all pitch extractors performed poorly on speaker LW (female) and well on speaker BF (also female). The low-pitched male speaker DG was a problem for H-S and AMDFD, but not so for others. Each pitch extractor has its own most favorite and least favorite speakers as illustrated in Fig. 6. Looking at the range of variation of the mean score, which is a measure of robustness over speakers, we find that AMDFD and Gold exhibit the smallest range, H-S exhibits the largest range, and ILS and JSRU exhibit a nearly equal range between these two extremes. We can make a similar set of comments on the plots shown in Fig. 7 for the noise condition. We observe that all pitch extractors performed substantially worse on speaker LW, with AMDFD failing severely. The high-pitched male speaker PH was a problem for Gold and H-S, but not so for others. The range of mean scores was large for all pitch extractors.

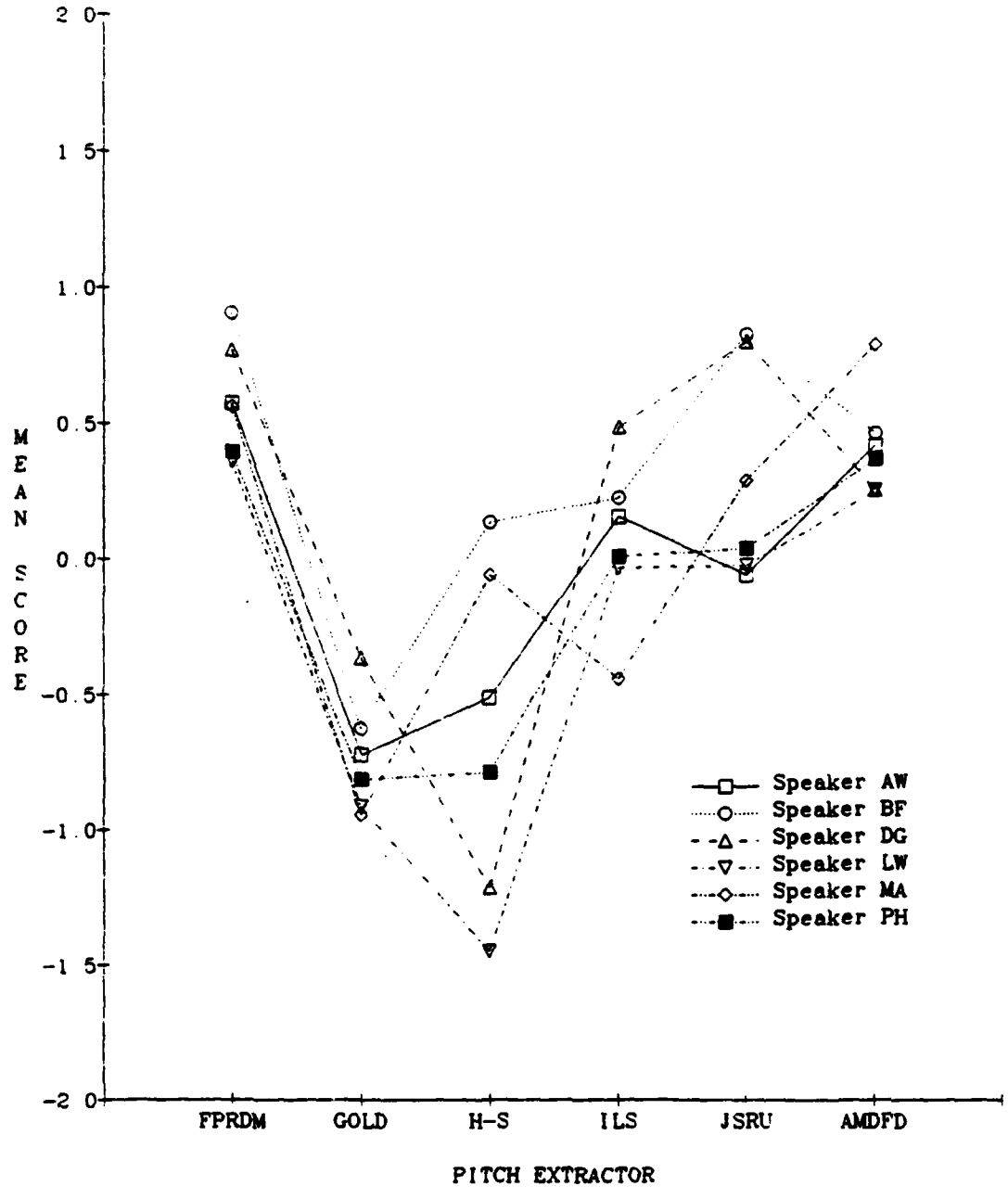



FIG. 6. Mean subjective scores for each speaker, under the clear condition.

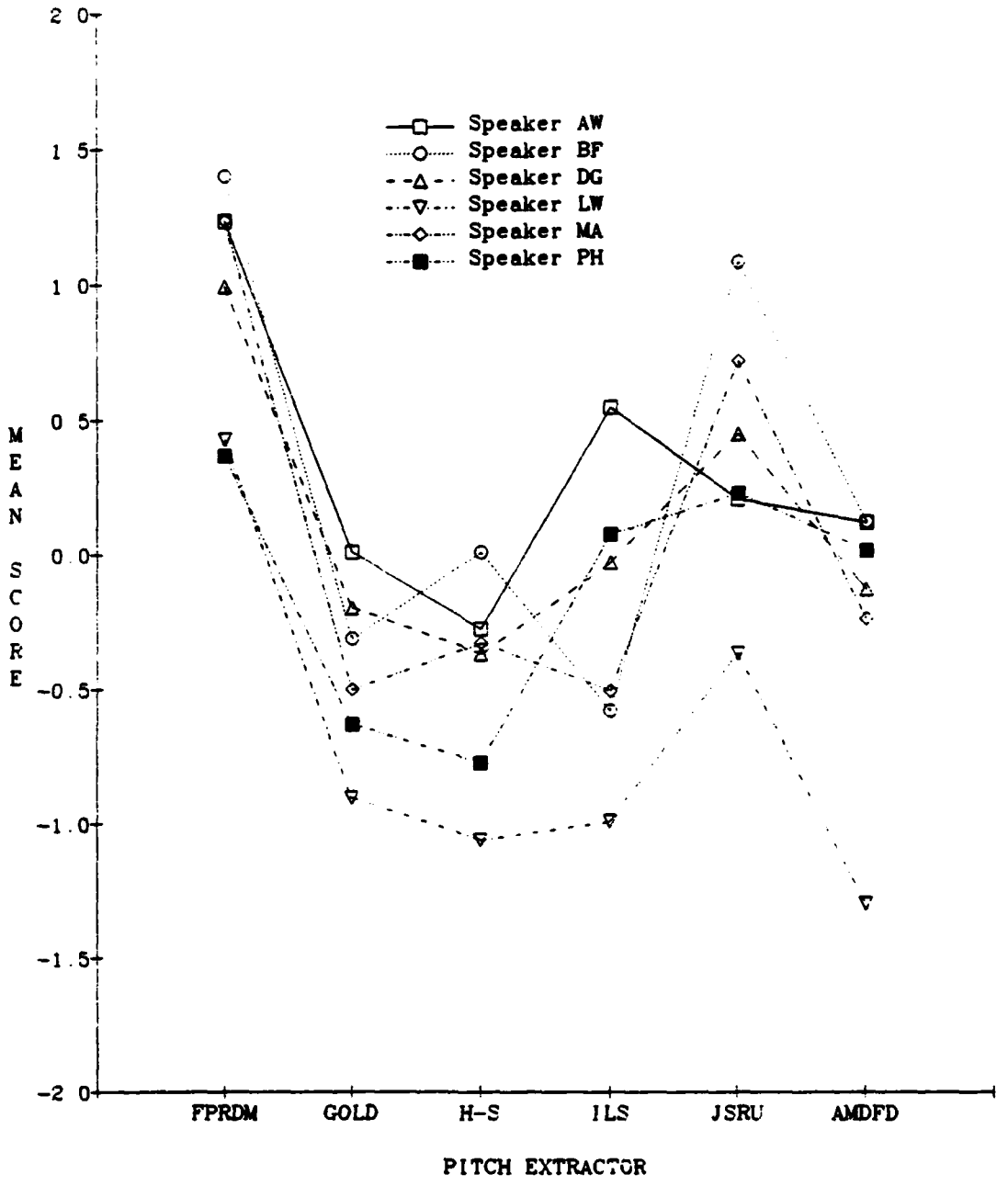



FIG. 7. Mean subjective scores for each speaker, under the ABCP noise condition.

Next, we consider the performance of pitch extractors as a function of the speech material. Figures 8 and 9 display the mean scores for each of the common sentences 1-6 given in Table 5 (we averaged over all six speakers and all subjects) for the clear and noise conditions, respectively. Each pitch extractor has its own most favorite and least favorite sentences. From Fig. 8, we see that the all-voiced sentence 1 produced good results for all but the H-S and AMDFD pitch extractors. The Gold pitch detector performed particularly poorly on sentences 4 and 6. The range of mean scores was smallest for AMDFD and largest for Gold. From Fig. 9, we note that AMDFD performed quite poorly on sentence 2. The range of mean scores was smallest for H-S and largest for AMDFD.

Since the results presented above show AMDFD and JSRU to be the two best pitch extractors, we have plotted the mean scores for them and FPRDM against the 48 speaker-sentence combinations in Figs. 10 and 11, for the clear and noise conditions, respectively. We see from Fig. 10 that AMDFD performed quite poorly on the sentence DG1, as also noted above, but otherwise AMDFD's ratings were generally better and varied over a narrower range as compared to JSRU's ratings. Also, we see several sentences over which FPRDM's rating was exceeded by the rating of either AMDFD or JSRU (e.g., DG5 and MA1). From Fig. 11, we readily see the inferior performance of AMDFD as well as its

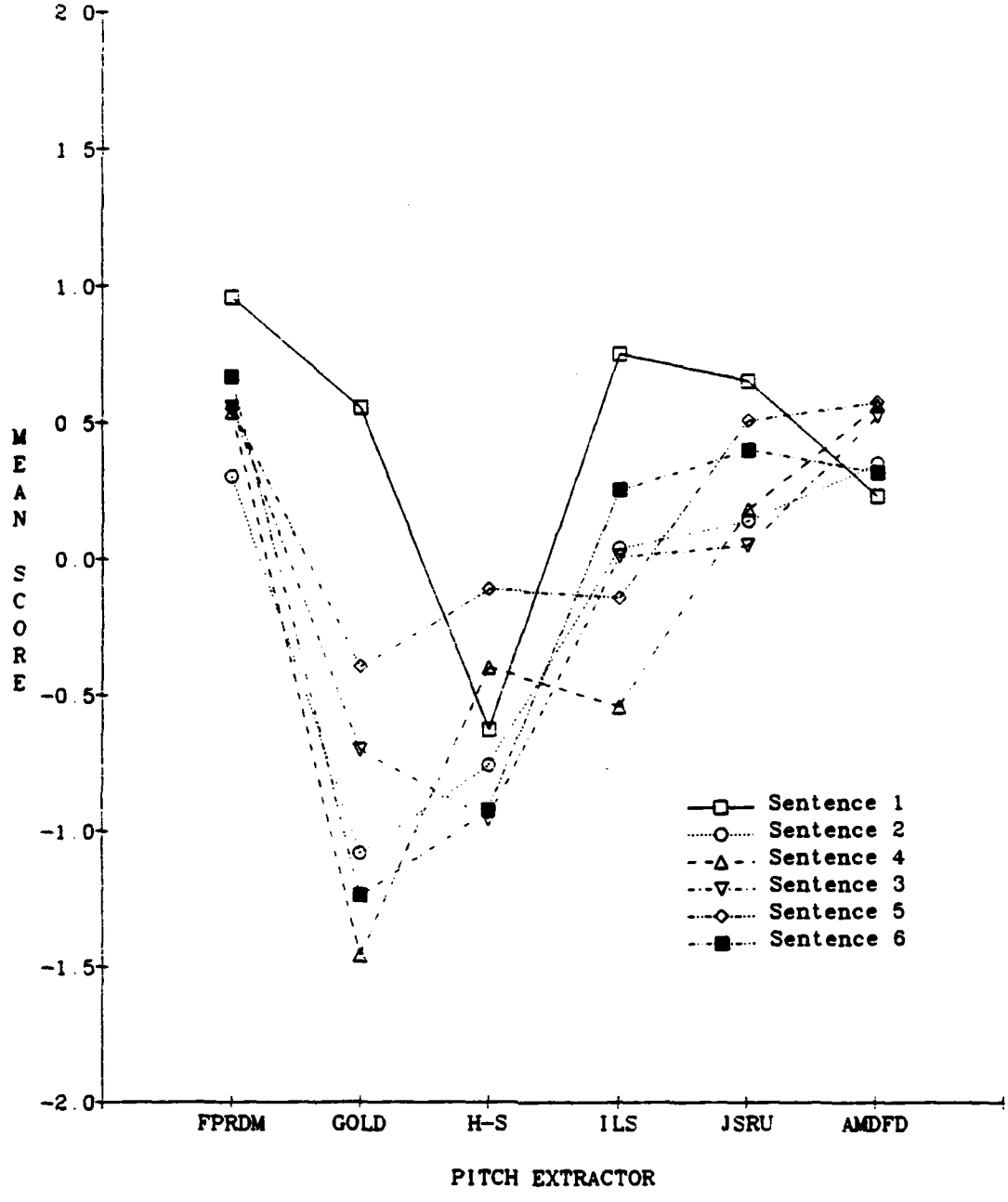



FIG. 8. *Mean subjective scores for each of the six common sentences, under the clear condition.*

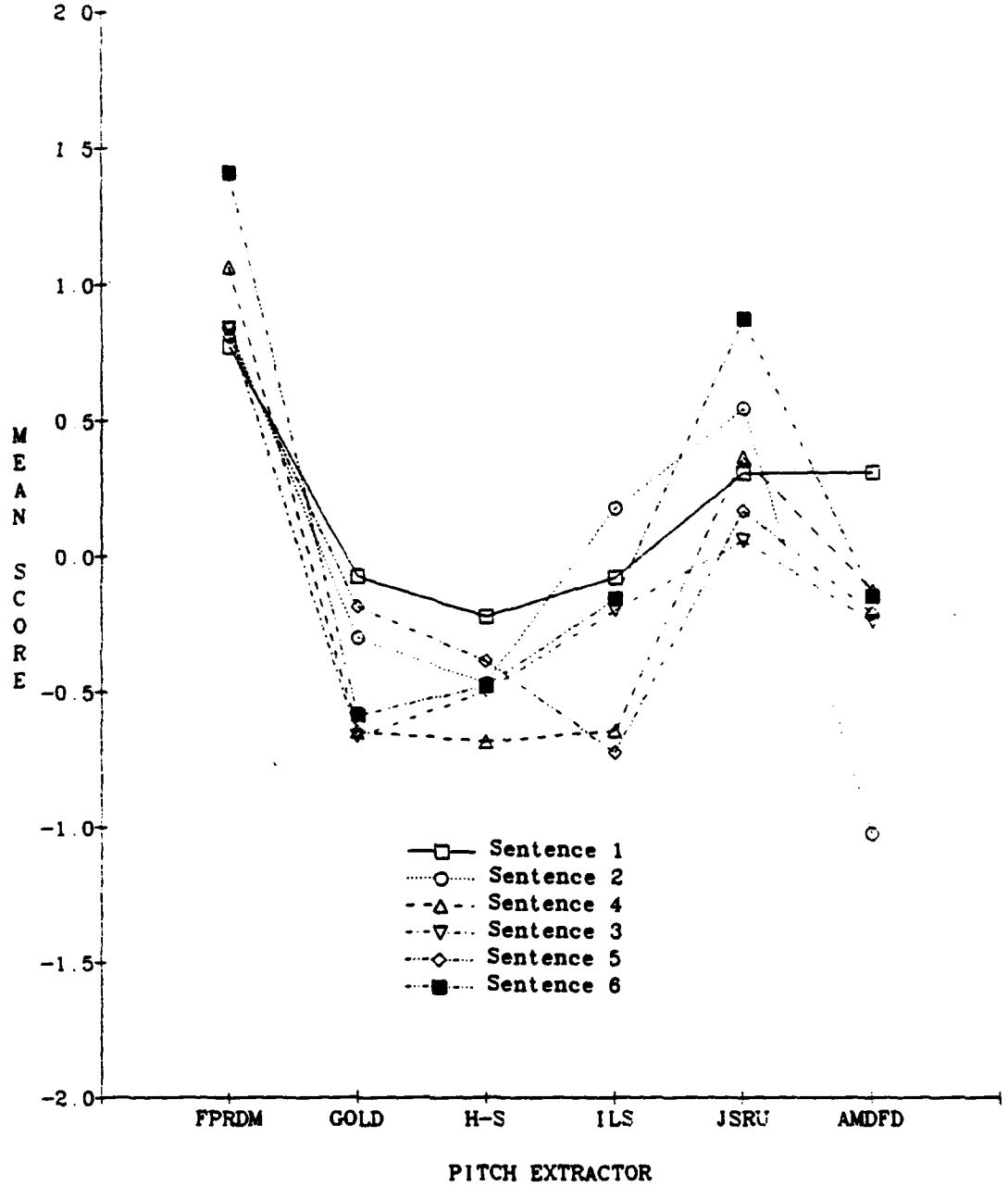



FIG. 9. Mean subjective scores for each of the six common sentences, under the ABCP noise condition.

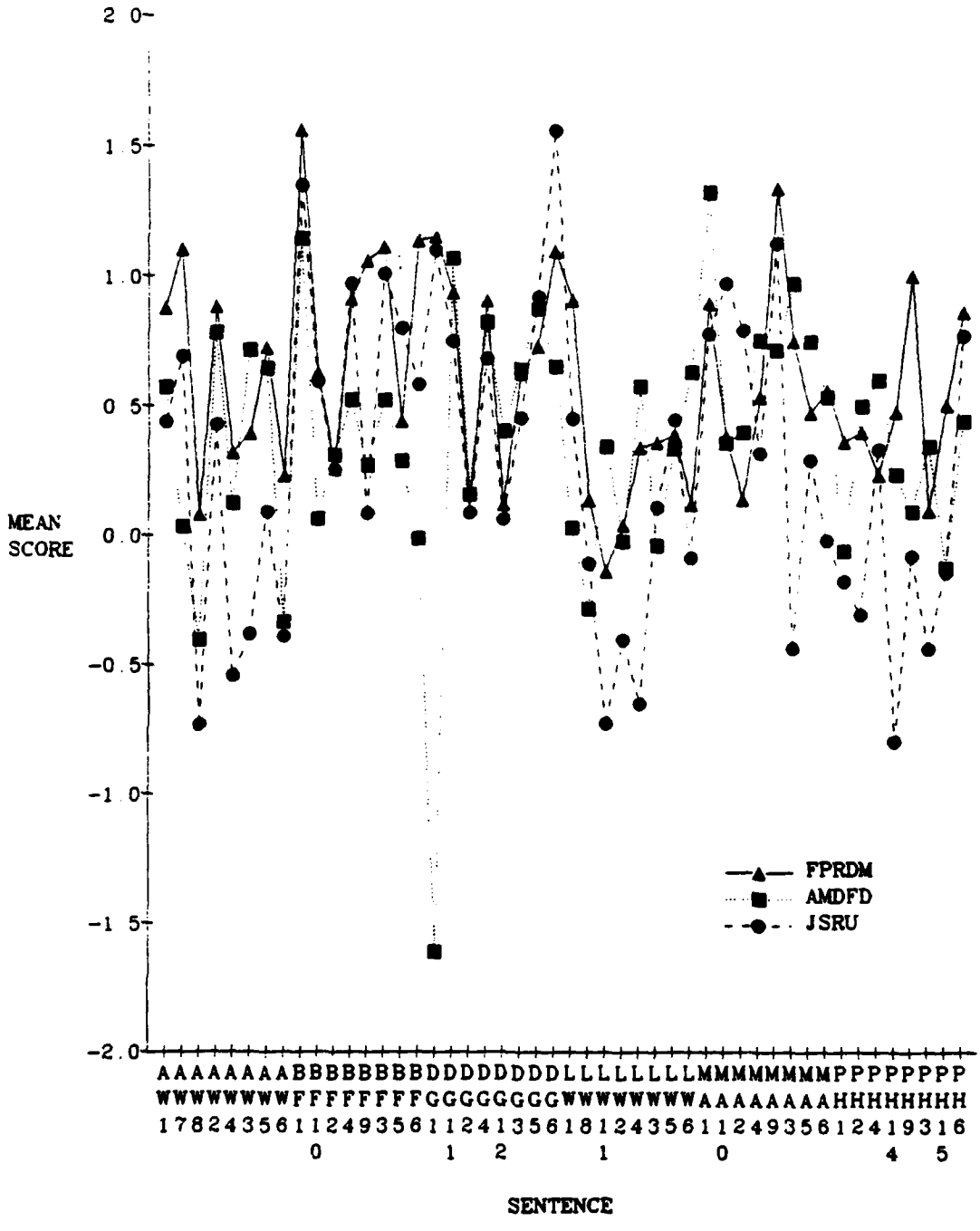



FIG. 10. Mean subjective score plotted as a function of the 48 speaker-sentence stimuli, for three pitch extractors under the clear condition.

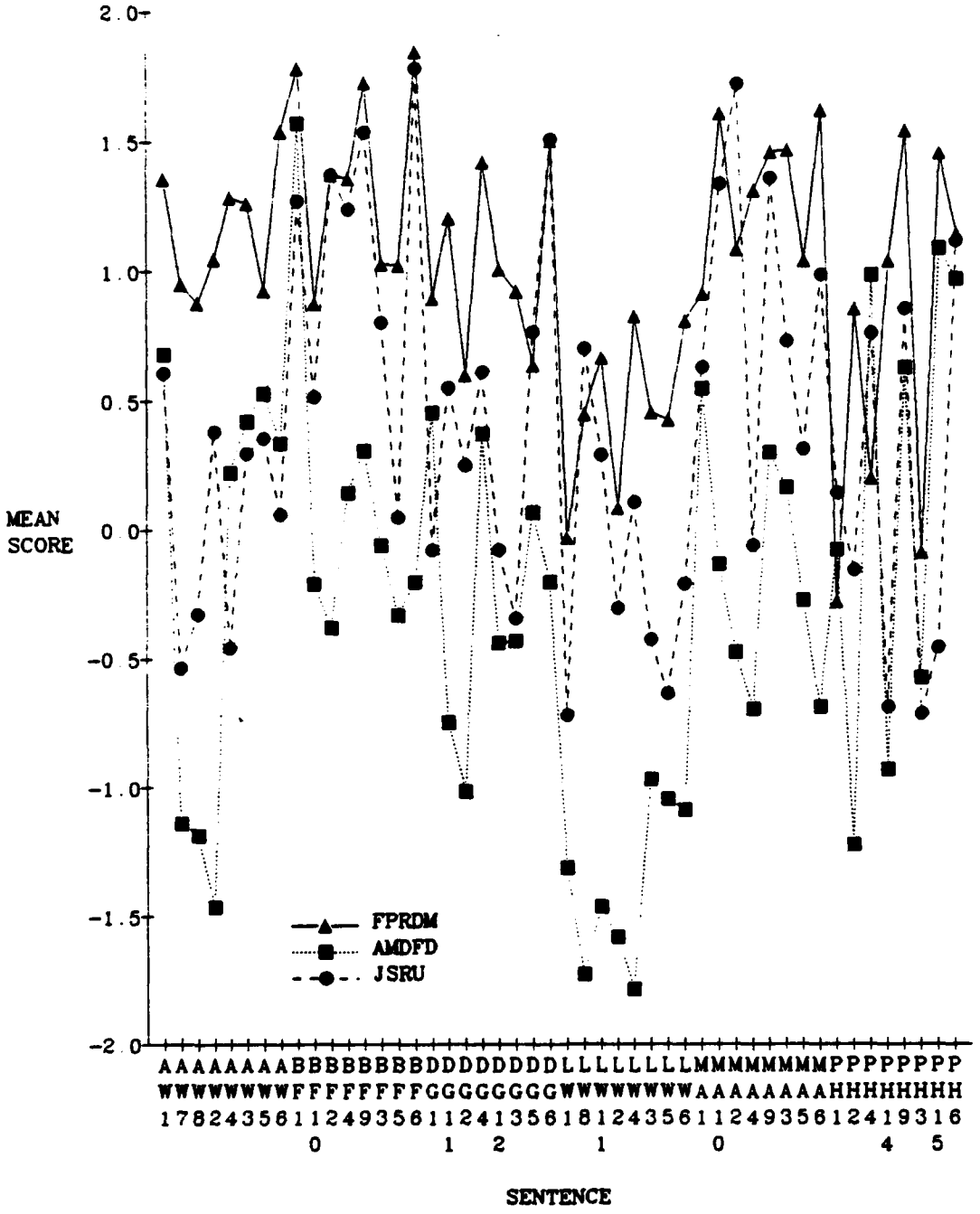



FIG. 11. Mean subjective score plotted as a function of the 48 speaker-sentence stimuli, for three pitch extractors under the ABCP noise condition.

performance variation over a wider range. AMDFD performed poorly on at least some of the sentences from all speakers but BF and MA, with its ratings being the worst for speaker LW. Again, we see a few cases in which FPRDM's rating was exceeded by JSRU's or AMDFD's rating.

Using a separate statistical package available on our DECSystem-20 computer, we performed two six-way analyses of variance, one for the clear condition and one for the noise condition (speakers x sentences x coders x pitch extractors x subjects x replications). We included in the analyses only the six common sentences. The results show that both speaker and sentence were highly significant sources of variance, but that the interaction of speaker and sentence was much more significant than either. This means that although there was some similarity among different sentences spoken by the same speaker, and among different speakers saying the same sentence, it is better to regard each speaker-sentence combination as unique. Therefore the analyses were repeated, replacing the 6 speakers x 6 sentences dimensions by a single dimension representing all 48 stimulus sentences. For both the clear and noise conditions, all main effects discussed above were significant, although the difference between the HDV and LPC coders only just reached significance ($P = 0.38$ for clear; $P = 0.49$ for noise). The effect of subjects was significant at about $P = 0.01$. All the other main effects, and all of the

interactions that included the stimulus sentences, were extremely significant ($P < 0.0001$).

Considering again the comparison between the AMDFD and JSRU pitch extractors, we restate the results that AMDFD performed better in the clear and JSRU performed substantially better in ABCP noise. While there is one clear condition (admittedly, "clear" is not unambiguous), there are a number of operational noise conditions. For example, the Department of Defense (DoD) typically evaluates speech coder performance over the noise conditions that include ABCP noise, office noise, ship noise, and tank noise. As we evaluated the pitch extractors only in ABCP noise, we do not have sufficient evidence to recommend JSRU over AMDFD for the DoD applications, for example. Also, we understand from a recent conversation with Tom Tremain of the DoD that version 45 of LPC-10 includes an improved AMDFD algorithm, which has produced better intelligibility scores in noise.

## 7. PERCEPTUAL EFFECTS OF PITCH AND VOICING ERRORS

Pitch extractors can and do produce several types of pitch and voicing errors at various locations within an utterance. To develop meaningful objective measures for evaluating pitch extractors, it is necessary to identify individual error types and patterns that create distinct perceptual effects and weight them according to their influence on speech intelligibility and quality. To identify and isolate the perceptual effects of individual error types, we developed a program to introduce in a controlled manner specific errors of known magnitude and duration at specified locations in the reference FPRDM pitch contour. We conducted an experimental study involving informal listening tests on the output speech of the LPC coder that used such perturbed pitch contours, to assess the perceptual effects of different types of pitch and voicing errors and thereby gain some insight for developing objective pitch evaluation measures. In Section 7.1, we describe the program we developed for generating perturbed pitch files. We present our experimental results on the perceptual effects of voicing errors in Section 7.2 and of pitch errors in Section 7.3.

## 7.1 Controlled Generation of Pitch and Voicing Errors

Our program for perturbing or corrupting the pitch in a controlled manner, called CORPTCH, has two parts. The first part collects characteristics about an input utterance. These characteristics, in conjunction with user-defined constraints, are used to specify regions of the utterance where errors are permitted. The second part of the program introduces pitch and voicing errors in these regions.

To obtain the required characteristics, the program reads the reference FPRDM pitch data from a file. The corresponding speech file is also accessed, and a frame-by-frame energy contour is computed. Voicing transitions are located, and statistics of the pitch and energy contour are obtained. Measures of the dynamics of the pitch and energy are also computed. The details of these steps are given below.

The relative locations of the voicing transitions are specified at each frame as the distance from the current frame to the last transition and to the next transition. The distances are defined as a percentage of the current region length and also as the number of frames. The distances are used to position errors relative to the location of voicing transitions.

The statistics that are computed are the mean, standard deviation, and median of the energy and pitch over the entire utterance. The median pitch for each voiced region is also evaluated. These measures are used to position the errors with respect to the realtive magnitude of the pitch and energy.

To determine the dynamics of the pitch or energy in the neighborhood of a given frame, we developed two measures: a "representative" slope, RS, which indicates the direction and rate of change of the parameter (pitch or energy) at each frame and a reliability factor, RF, which is a measure of the accuracy of the slope. RF also indicates whether the contour is smooth or "noisy" in the region. To obtain the two measures for either pitch or energy, the parameter contour is 3-point median smoothed to remove outliers. At each frame a difference is computed between the next frame smoothed value and the last frame smoothed value. Notice that this difference (referred to as smoothed difference below) represents the trend of the parameter over three smoothed frames or over five unsmoothed frames. However, this measure is not a good indicator of the slope if the parameter contour is noisy. To obtain information about the smoothness of the contour, two more differences (referred to as the unsmoothed differences below) are computed: the absolute value of the difference between the last frame and current frame unsmoothed values and the absolute value of the difference between the next frame and the

current frame unsmoothed values. The sum of the two differences is the total change in the parameter over the three-frame interval. The reliability factor RF is then defined as the absolute value of the ratio of the smoothed difference to the sum of the unsmoothed differences. RF is a positive fraction with a maximum value of one. An RF of unity implies that the smoothed difference is the actual slope of the parameter at the current frame and that the contour is smoothly varying. When either the numerator or the denominator is zero, RF is computed as follows. If the sum of the unsmoothed differences is zero, then the smoothed difference will also be zero indicating that the parameter is not changing. RF is set to unity. If the smoothed difference is zero and the unsmoothed difference sum is not, a peak or null in the contour has occurred. To indicate the sharpness of the transition, RF is computed as the reciprocal of the unsmoothed difference. Once RF is computed, the representative slope RS is computed as the square of RF times the smoothed difference.

After all the characteristics have been computed, errors are introduced in the pitch data. The user specifies the errors to be introduced. The program permits a number of error categories including gross, fine (or jitter), doubling, halving, VUV, UVV, a shift of the average fundamental, and a change in the variance of the pitch contour. For each error category, the

user specifies one or more thresholds and parameters that control the intensity, frequency, and relative location of the error. At each frame, the pitch, the energy, the two measures of dynamics (RF and RS) for both pitch and energy, and the transition location measures are compared against respective thresholds. If all the quantities are within the user-defined bounds then a pitch or voicing error is permitted. Other user-defined parameters, such as the number of errors, are examined to determine if an error must actually occur at the current frame.

In general, the pitch is corrupted using a single error category. However, if more than one error category is chosen, the errors are created in the following order. First, voicing errors are made at all designated frames in the utterance. These frames are not modified by any subsequent processing. Second, fine errors are introduced in all chosen frames in the utterance. Finally, gross errors including pitch doubling and halving are introduced. A gross error overrides a previously defined fine error. If mean and variance changes in addition to one or more of the above error categories are specified, a second-pass processing of the pitch file is required.

We conducted a number of experiments to examine the perceptual effects of specific types of pitch and voicing errors. For each experiment, we produced

a set of perturbed pitch files with our CORPTCH program, by introducing selected types of error at known locations in the reference FPRDM pitch contour. The perceptual effects of these errors on speech intelligibility and quality were then evaluated by informal listening tests of synthesis that used the perturbed pitch files. We used the LPC coder for generating the synthesized speech. For many of the experiments, only a small subset of utterances from our speech database were tested. Consequently, the results of these experiments, reported below, should not be regarded as conclusive, but rather as indicators of the types of properties and characteristics that pitch and voicing errors can exhibit.

## 7.2 Perceptual Effects of Voicing Errors

We conducted several experiments to assess the perceptual effects of voicing errors. Our first experiment was designed to examine the effects of VUV and UVV errors at transitions. These errors are quite common, since correct determination of the voicing state at transitions is a difficult task for most pitch extractors. For the test we used six utterances, one sentence spoken by all six speakers. The sentence, "A thickset officer pitched out her hash", was chosen since it contained a number of transitions. For each of the six sentences, four perturbed pitch files were created, each containing one of

the following error types: (1) a VUV error at the first frame of each voiced region, (2) a VUV error at the last frame of each voiced region, (3) an UVV error at the first frame of each unvoiced region, and (4) an UVV error at the last frame of each unvoiced region. The pitch frequency of the nearest voiced frame was inserted in frames containing UVV errors.

Informal listening tests confirmed that voicing errors of just a single frame in duration can cause noticeable distortions in the speech. The speech with VUV errors lacked clarity and crispness and was characterized as being choppy, raspy, and noisy. The presence of UVV errors caused the speech to sound slurred and buzzy. Some tonal or ringing effects were also noted in these utterances. Listeners (we used up to three experienced listeners) invariably found that UVV errors were not as objectionable as the VUV errors. VUV errors at the beginning of voiced regions generally appeared to degrade the speech intelligibility and quality more than those at the end of voiced regions. Changing of essential perceptual cues at the onset of voiced regions because of the presence of voicing errors could be responsible for this result. A similar result was obtained when we compared the perceptual effect of UVV errors at the beginning of unvoiced regions with that at the end of unvoiced regions, although the difference in this case was less severe. Judgments on the severity of the foregoing four types of voicing errors and

their dependency on location differed greatly across the various utterances and among the listeners. These differences might have been caused to some extent by listener preference but largely because of the difference in the energy of frames at which the errors occurred. The average energies of the first and the last frames of unvoiced regions were approximately equal at 28.5 and 27.2 dB, respectively, whereas the average energies of the first and the last frames of voiced regions were higher at 42.7 and 35.2 dB, respectively, and differed by 7.5 dB.

A second experiment was conducted to examine the influence of frame energies on the perceptual effects of voicing errors. VUV errors were tested since the frames where they occurred contained the largest variation in energy. Two sets of perturbed pitch files were created. One set contained errors at the first frame of each voiced region only when the frame energy was above a threshold of 43 dB. The other set contained errors at the first frame of each voiced region only when the energy was below the same threshold. Similar test data was also generated for VUV errors at the end of voiced regions using an energy threshold of 35 dB. The thresholds used resulted in approximately an equal number of errors in both the high and low energy regions for most sentences. Informal listening tests indicated that errors at high-energy frames always caused more adverse effects than those at low-energy

frames. From listening tests of various parts of sentences from the first experiment described above, we were able to conclude that the severity of UVV errors at transitions was also dependent on frame energies.

We conducted a third experiment to determine the effect of error location and type when energy was not a factor. A single sentence was chosen that contained several transitions with approximately the same energy levels and where the energy changes across the transitions were small (a total of 7 transitions). VUV and UVV errors of a single frame duration at these transitions were compared. The difference in the severity of the two error types or the perceptual difference caused by location was rather small. These observations seem to indicate that energy rather than error type or location is the important factor. However, since only a single utterance was tested, it is difficult to draw any sound conclusions from the test.

In all of the above experiments, only a single frame at each transition contained a voicing error. We also examined the effect of VUV errors of two frames in duration at each transition for the six-sentence database. The resulting speech had a whispered quality that was quite obvious. This observation indicates that the duration of the voicing error is an important factor.

Voicing errors in the middle of voiced and unvoiced regions were also examined. Three test sentences were chosen that contained sections of voiced regions where the energy dropped substantially for short intervals. Pitch extractors have a tendency to cause VUV errors in these regions. Listening tests indicated that errors in these regions of a single frame in duration had little or no effect on speech quality or intelligibility. However, if two or more consecutive errors occurred, they did cause perceivable raspiness. From the previous experiments it was clear that VUV errors at high-energy frames would cause substantial degradation, so no further tests on these error types were conducted.

To examine the effect of UVV errors in the middle of unvoiced regions, a single sentence was chosen; this sentence contained several unvoiced regions that differed substantially in energy levels. Five perturbed pitch files were created, each containing UVV errors in a different energy region. The five energy regions were: below 10 dB, 10 to 20 dB, 20 to 30 dB, 30 to 40 dB, and above 40 dB. Errors of a single frame in duration in regions below 30 dB had minimal effect on the speech. However, consecutive errors of 2 or more frames caused the speech to be buzzy. All errors in the high-energy regions (above 30 dB) produced audible effects. Errors in the highest-energy region caused severe distortions that reduced the intelligibility of the speech.

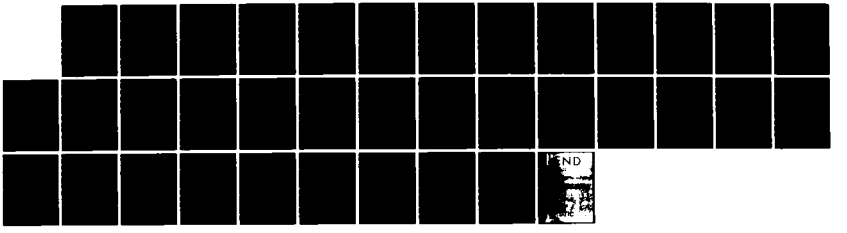
We summarize the results of our listening tests on voicing errors as follows:

- The effect of voicing errors (VUV or UVV) is highly dependent on frame energies; the higher the energy of the frame at which an error occurs the larger the perceived distortion.
- Errors that occur in two or more consecutive frames are much more audible than isolated frame errors.
- VUV errors cause raspy, noisy effects, whereas UVV errors cause slurring and buzziness. UVV errors are in general not as objectionable as VUV errors.
- Some evidence exists that errors at the beginning of voiced regions cause more adverse effects than those at the end of voiced regions.

## 7.3 Perceptual Effects of Pitch Errors

We conducted several experiments to assess the perceptual effects of pitch errors. From the different error measures and other output obtained using the program PEVAL for the various pitch extractors (see Section 4.3), we observed that many of the gross pitch errors occurred at the beginning and end of voiced regions. Large variations in pitch can occur in these regions, and most pitch extractors have a tendency to smooth these variations, which produces poor estimates of the actual pitch dynamics. To examine these issues, we chose twelve utterances, two sentences spoken by the six speakers, and manually changed the FPRDM pitch values of up to three frames at the

AD-A146 838 SUBJECTIVE AND OBJECTIVE EVALUATION OF PITCH EXTRACTORS 2/2
FOR LPC AND HARMO..(U) BOLT BERANEK AND NEWMAN INC
CAMBRIDGE MA V R VISWANATHAN ET AL. JUL 84 BBN-5726
UNCLASSIFIED DCA100-83-C-0023 F/G 20/1 NL



beginning and end of voiced regions. To assure that only highly dynamic regions were modified, only pitch values with a representative slope (RS) of greater than 5% were changed. The pitch values were altered so as to yield a flat pitch contour. A substantial number of the pitch changes resulted in errors of 10 to 20%, and some errors were as high as 40 to 80%. Informal listening tests indicated that many of the errors caused little or no effect on speech quality or intelligibility. Upon comparing these utterances with synthesis without pitch errors, we did find a few regions that sounded slightly monotone. In these regions, the pitch contour had a large and consistent slope, and we had changed all three pitch values to yield a flat local pitch contour, thus altering the natural trend of the pitch contour. These observations were surprising, considering the results we obtained for voicing errors in the same regions.

Smoothing of rapid changes in pitch over short intervals (2 to 5 frames) in the middle of voiced regions was also examined. Results indicated that these errors also had little or no effect on speech quality or intelligibility. These observations indicate that smoothing of the pitch contours (both reference and test) prior to objective measurement computations may remove errors that are not perceptually significant and could result in improved objective scores.

Another experiment was conducted using six sentences to examine the effects of pitch errors in smooth or non-noisy regions of the pitch contour. Errors of $\pm$4% were inserted in regions where the pitch contour was flat and non-noisy. For each of the utterances, an equal number of errors were also made in regions where the pitch contour was smoothly increasing or decreasing. Similar test data with errors of $\pm$8% were also generated. Changes in speech quality were noted in all utterances containing errors in the flat regions. The speech was characterized as being more bubbly. The flat region errors of $\pm$8% caused noticeable distortions in the dynamics of the pitch that were not natural. Pitch errors of $\pm$4% in regions of increasing or decreasing slope were not detectable, whereas errors of $\pm$8% caused some noticeable but not unnatural changes in the pitch dynamics. In the latter case, the perturbed pitch and the reference pitch produced natural-sounding speech, and there was no clear preference between the two. The perceptual effects of errors in both the flat and sloped regions were more pronounced for the female speakers and when the errors occurred in several consecutive frames.

We summarize the results on pitch errors as follows:

- Large pitch errors at the beginning or end of voiced regions may not be perceptually relevant if they do not disrupt the natural trend of the pitch contour over several frames.

- Smoothing of short-duration rapid changes in the pitch contour in the middle of voiced regions may not degrade the speech quality and intelligibility. This suggests smoothing of the test and reference pitch contours before they are compared for objective evaluation.
- The adverse effects of pitch errors are dependent on the magnitude of the errors, the magnitude of the pitch frequency, the duration of errors, and the total number of gross errors.
- Pitch errors in regions where the pitch contour is flat are more noticeable than those in highly dynamic regions.

Again, we wish to stress that the results presented above should be viewed as empirical and should provide only general indications of perceptual relevance (as viewed through LPC synthesis) of pitch and voicing errors. Objective measures that incorporate one or more of the properties presented above are described in the next chapter.

# 8. OBJECTIVE EVALUATION OF PITCH EXTRACTORS

In this chapter, we describe a large number of objective measures we developed and investigated for the evaluation of pitch extractors (Section 8.1); present the results we obtained by correlating the objective scores with the mean subjective rating scores from our formal subjective tests described in Chapter 6 (Section 8.2); and recommend a set of objective measures each of which produced consistently high correlation for both LPC and HDV coders, under both clear and ABCP noise conditions, and in two evaluation conditions, one involving the complete database of 48 stimulus sentences (six speakers x eight sentences) and one involving eight subsets of six stimulus sentences each (Section 8.3).

## 8.1 Development of Objective Measures

In Subsection 8.1.1, we briefly review the basic objective pitch and voicing error measures and present the results we obtained using these basic error measures over the 48-sentence speech database. Subsection 8.1.2 describes several methods of weighting the basic error measures, which we developed using the results of our perceptual study given in Chapter 7. In Subsection 8.1.3, we outline the procedure that we used in computing a large number of objective measures.

### 8.1.1 Basic Error Measures

As we mentioned in Section 4.2, there are three situations in which a pitch extractor under test causes an error to occur; in terms of the voicing status of the true or reference pitch and the test pitch, respectively, these three situations are denoted as VUV, UVV, and VV. The basic error measure, given in Section 4.2, for either VUV or UVV error type is the number of frames containing the respective error type expressed as a percentage of the total number of frames of data used in the evaluation. The VV case involves two types of error, gross pitch error and fine pitch error; the associated basic error measures are percent gross pitch error (i.e., percentage of the frames containing gross pitch errors), fine pitch error mean, and fine pitch error standard deviation. In our objective evaluation work, we used primarily the three basic error measures: the percent VUV error, the percent UVV error, and the percent gross pitch error, and the total error measure, which is the sum of these three basic error measures. We have referred to these measures as basic measures since they do not involve any form of weighting based on such quantities as speech signal energy.

Before we describe methods of weighting the foregoing basic error measures, we present the results we obtained using the program PEVAL and the

basic error measures over the 48-sentence speech database used in our formal subjective tests. The error results for the five pitch extractors under evaluation are given in Table 6 for the clear case and in Table 7 for the ABCP noise case. In computing the error results, we used a frame size of 20 ms (or a frame rate of 50 frames/s) to correspond to the frame size used by both the LPC and HDV coders, which provided the test stimuli. Pitch frequency doubling and halving errors, which are included in the gross pitch error, are also given in the tables as percentages.

| Error | AMDFD | Gold | H-S | ILS | JSRU |
|---|---|---|---|---|---|
| Percent VUV Error | 7.31 | 12.64 | 13.56 | 13.38 | 7.33 |
| Percent UVV Error | 2.25 | 6.61 | 1.04 | 0.73 | 2.50 |
| Percent Gross Pitch Error | 4.66 | 4.52 | 8.87 | 2.44 | 1.90 |
| Total Error | 14.22 | 23.77 | 23.47 | 16.55 | 11.73 |
| Pitch Doubling | 0.25 | 0.08 | 1.32 | 0.00 | 0.00 |
| Pitch Halving | 0.05 | 0.60 | 0.67 | 0.55 | 0.50 |

TABLE 6. Basic pitch and voicing error results for the five pitch extractors, computed over the 48-sentence clean-speech database.

| Error | AMDFD | Gold | H-S | ILS | JSRU |
|---|---|---|---|---|---|
| Percent VUV Error | 18.28 | 29.61 | 7.15 | 32.79 | 26.32 |
| Percent UVV Error | 11.42 | 2.14 | 15.52 | 1.89 | 0.20 |
| Percent Gross Pitch Error | 12.79 | 4.79 | 14.83 | 3.62 | 0.93 |
| Total Error | 42.49 | 36.54 | 37.51 | 38.30 | 27.45 |
| Pitch Doubling | 0.08 | 0.03 | 1.39 | 0.0 | 0.0 |
| Pitch Halving | 6.59 | 2.46 | 3.65 | 1.89 | 0.25 |

TABLE 7. Basic pitch and voicing error results for the five pitch extractors, computed over the 48-sentence ABCP noise-added speech database.

From Table 6, we note that the voicing error results for AMDFD and JSRU are about the same and are considerably superior to those for the other three pitch extractors. The two cepstral pitch extractors, ILS and JSRU, produced the lowest gross pitch error. The JSRU algorithm also produced the lowest total error. The rank ordering of the pitch extractors using the total error as the criterion corresponds to the rank ordering by the mean subjective rating score, with the one exception that the total error reverses the order of the top two pitch extractors.

Table 7 shows that gross pitch error and total voicing error increased substantially in noise, for all five pitch extractors. Because of the added ABCP noise, a substantially large number of frames were declared unvoiced resulting in a large VUV error for all but the H-S pitch extractors. While one would expect to see some reduction in the UVV error for the same reason, both AMDFD and H-S produced substantially higher UVV error in noise than in clear. It is reasonable to assume that the higher UVV error was also responsible for the higher gross pitch error for these two pitch extractors. As in the clear case, the JSRU method produced the lowest gross error and the lowest total error. Considering the rank ordering of the pitch extractors, the total error is in agreement with the mean subjective score only as far as the best pitch extractor is concerned.

From a comparison of the results given in Table 3 for the six TI sentences (see Section 4.3) with those given in Table 6 for our 48-sentence database, we find that the total error given in Table 6 is substantially larger than that given in Table 3 for Gold and H-S. A detailed examination uncovered the fact that we had inadvertently used an incorrect time delay of 60 ms (or three 20-ms frames) for the Gold pitch detector. We recomputed the total error over the 48-sentence database, using different values of time delay as discussed in Section 4.3. A delay of 20 ms produced a total error of

13.2%, and a delay of 40 ms produced a total error of 17.2%. Had we used a 20-ms delay, the Gold pitch detector might have yielded substantially better subjective rating scores than we reported in Section 6.4. To be consistent with the already gathered subjective data, however, we continued to use the 60-ms delay for the Gold pitch detector in our subsequent objective evaluation work. A similar recomputation of the total error for the H-S method indicated that we had used the correct or minimum-error delay of 20 ms for this pitch extractor.

### 8.1.2 Methods for Weighting the Errors

As mentioned above, we considered in our objective evaluation study three types of errors: VUV error, UVV error, and gross pitch error. The basic or unweighted error measure for each error type assigns a value of one to each occurrence of the error, computes the total value over the database, and normalizes it by dividing with the total number of frames and multiplying with 100 (to get the result in percentage). The idea of weighting each frame error is to reflect the perceptual significance of the error in some manner, with the expectation that the weighted error over the database produces a higher correlation with the subjective rating scores than does the unweighted error.

From the experimental results we reported in Chapter 7 on the perceptual

effects of pitch and voicing errors, we chose to investigate four methods of weighting: weighting based on the speech signal energy over individual frames, weighting based on the duration of consecutive errors, weighting based on the pitch frequency or the magnitude of the pitch frequency error, and weighting that accounts for the context in which the error occurs. For each weighting method, we conducted several initial tests to determine one or more appropriate forms for the weight; in these tests, we computed the correlation of the weighted error with the mean subjective scores (see Section 8.2) for the evaluation of the different forms we considered for the weight. Below, we describe the four weighting methods and indicate the form(s) we chose for the weight, in each case.

The importance of energy weighting is clear from the results of our perceptual study reported in Chapter 7. Our perturbation experiments showed that pitch and voicing errors occurring in high-energy frames produced more audible effects in the synthesized speech than those in low-energy frames did. To emphasize errors at high-energy frames, we considered three forms for the energy weight: the RMS value of the speech signal over a frame, the RMS value in decibels, and the RMS value divided by the maximum frame RMS value over the individual test utterance. Reference [10] uses the third form.

From the results of our perceptual study of pitch and voicing errors, we note that errors occurring in two or more consecutive frames are substantially more audible than are isolated frame errors. To account for this duration effect, we chose empirically the following weighting method: a weighting factor of unity for isolated frame errors, a weighting factor of 1.5 for a duration of two to five frames, and a weighting factor of 2.0 for a duration of six or more frames.

For pitch-frequency weighting, the weighting factor we considered is FR/FMAX for both VUV and gross pitch errors and FT/FMAX for UVV errors, where FR is the reference pitch frequency, FT is the test pitch frequency, and FMAX is the maximum permissible pitch frequency. We used FMAX=500 Hz. For pitch-error weighting, the weighting factor we considered is $(|\ FT-FR\ |/FR)^r$; we used r=1. In our investigation, we used the above pitch-frequency weighting for VUV and UVV errors and the above pitch-error weighting for gross pitch errors.

The results of our perceptual study show clearly the perceptual importance of context (or location) in which pitch and voicing errors occur. There are three parts in our implementation of the context-dependent weighting function. The first part, which was motivated by the observed perceptual

significance of the voicing errors at unvoiced-voiced transitions (see Section 7.2), is to penalize the early and the late start of voicing in these transitions. Empirically, we chose a weight factor of 2 for the VUV errors occurring in the first three frames of the voiced region and for the UVV errors occurring in the last three frames of the unvoiced region, in any unvoiced-voiced transition; we used a unity weight for all other voicing errors. Second, we recall from Section 7.2 that large pitch errors at the beginning and at the end of a voiced region and large pitch errors caused by smoothing of short-duration rapid changes in the true pitch in the middle of a voiced region did not produce significant changes in perceived speech quality and intelligibility. We chose a weight of 0.1 for gross pitch errors occurring in the first two and the last two frames of a voiced region, provided that the local slope RS is greater than 10. We chose the same weight for gross pitch errors in the middle of a voiced region, provided that the reference pitch contour is noisy (RF $\leq$ 0.5 and RS $\leq$ 10). All other gross pitch errors were assigned a weight of unity. The third part deals with the threshold used in deciding if a pitch error is a gross pitch error or not. The nominal value used for the threshold is 10%. In voiced regions where the reference pitch contour is flat (RS $\leq$ 4) and non-noisy (RF $\geq$ 0.8), we lowered the threshold to 5%, to account for the observed increase in listener's sensitivity to pitch errors in such regions (see Section 7.2). We also note

that the difference limen (or just-noticeable difference) for pitch frequency is 0.3% to 0.6% for a flat (monotone) pitch contour [22] and about 2% for a linear (ramp) pitch contour [23]. We reiterate that the presence of the context included in the foregoing three parts is determined from the reference pitch data.

Since we considered the use of no weighting or one or more of the four weighting methods for each of the three types of errors (VUV, UVV, and gross pitch errors), we had a total of 125 (5x5x5) possible combinations we investigated. As error measures, we considered each of the three types of errors separately, sum of any two types of errors, and sum of all three types of errors. This led to a total of 215 (3x5 + 3x25 + 1x125) error measures we investigated. The total of all three types of errors produced, in general, higher correlation with subjective scores than did the one-at-a-time and the two-at-a-time error measures (see Section 8.2). It is convenient to use a simple notation to refer to the 125 total error measures. Let us use the order VUV error, UVV error, and gross pitch error in specifying the weights. Also, let us denote the unweighted case by the letter C (for count); the energy weighting by the letter E; the duration-based weighting by the letter D; the pitch frequency and pitch error weighting by the letter F (for frequency); and the context-dependent weighting by the letter L (for

location). Thus, the notation C-C-C refers to the total unweighted error measure, and the notation EDFL-C-EF refers to the total error measure that uses all four weighting methods for VUV errors, no weighting for UVV errors, and energy and pitch-error weighting for gross pitch errors.

We also implemented the objective measure used in [10]. We refer to this measure as the TI measure. This measure is a total of weighted VUV errors, weighted UVV errors, and weighted pitch errors (gross and fine pitch errors included). All errors are energy weighted using the factor (RMS value/maximum RMS) discussed above. In addition, pitch errors are weighted with the factors $[(FT-FR)/FR]^2$ and FR/500. Voicing errors in a frame at any voicing transition are weighted with a factor F/500 and all other voicing errors are weighted with a larger factor (1 + F/500), where F=FR for VUV errors and F=FT for UVV errors. This last-mentioned context-dependent weighting is contrary to our weighting method, which emphasizes the voicing errors at the unvoiced-voiced transitions.

### 8.1.3 Computation of Objective Measures

The procedure we used for computing the large number of objective measures described above was incorporated as part of the program PEVAL (see Section 4.2). The procedure is as follows. A frame-by-frame comparison is

made of the pitch data from reference and test pitch files over the speech database of interest, at a rate of 50 frames/s. Unprocessed speech waveform files are used to compute the frame speech energy required for energy weighting. Reference pitch data is used to locate the voicing transitions and compute each frame the slope RS and the reliability factor RF as discussed in Section 7.1; these are required for context-dependent weighting. For each frame that contains a pitch or voicing error, a set of weighting factors and products of these factors for different combinations of the weighting methods are computed. These weights and products of weights are summed over a database of pitch files, separately for each of the three types of errors (VUV, UVV, and gross pitch errors), normalized by dividing with the total number of frames processed for the cases involving no energy weighting and with the total of the frame energy factors for the cases involving energy weighting, and multiplied with 100 to obtain percentages. These sums are actually error measures since the unweighted value assigned to an occurrence of any error is unity. Composite error measures are then computed by adding any two of the VUV, UVV, and gross pitch error measures and by adding all three. In one session, the user computes the various objective measures for each of several pitch extractors; the computed objective measure data are all stored in one disk file to be used in subsequent correlation study. We produced two objective measure files, one for clean speech and one for ABCP

noise-added speech. Notice that the same objective measure file applies to both LPC and HDV coders, since they use the same pitch files as input.

## 8.2 Correlation with Subjective Rating

To evaluate and choose some good ones from the large number of objective measures we considered, we correlated the data from each objective measure against the mean subjective rating scores. We performed the correlation study in each of eight different conditions described below. First, we considered the overall scores for each of the four cases: LPC/Clear, HDV/Clear, LPC/Noise, and HDV/Noise. The overall subjective scores were obtained by computing, for each of the five test pitch extractors, the mean of 672 ratings for the clear condition (48 stimulus sentences x 7 subjects x 2 judgments) and the mean of 480 ratings for the noise condition (48 stimulus sentences x 5 subjects x 2 judgments). The overall objective scores were computed over the 48 sentences, once for the clear condition and once for the noise condition. (Recall that the objective scores are the same for both coders.) We had thus five objective scores and five subjective scores corresponding to the five pitch extractors, and we computed the correlation between the two sets of scores. We shall refer to this correlation as the 5-item correlation as it involves five scores. Second, we considered the scores at a more detailed

level, again for each of the four coder/background cases. We divided the 48 stimulus sentences into eight sets of six sentences each. For each of the first six sentences in Table 5, the set contained the same sentence spoken by all six speakers. From the remaining 12 stimulus sentences, we formed two sets by grouping together phonetically similar sentences. We considered the evaluation of each pitch extractor over each of the eight six-sentence sets, which produced 40 items to correlate over. We obtained the subjective scores by computing the mean of 84 ratings for the clear condition (6 stimulus sentences x 7 subjects x 2 judgments) and the mean of 60 ratings for the noise condition (6 stimulus sentences x 5 subjects x 2 judgments). We computed the objective scores over each of the eight six-sentence sets, once for the clear condition and once for the noise condition. We shall refer to the correlation for the second case involving detailed scores as the 40-item correlation. The 40-item correlation should in general be lower than the 5-item correlation. All correlations were computed using PEVAL under a separate command called CORRELATE.

A good, robust objective measure must produce high correlation under all eight evaluation conditions described above. In our investigation, we required high correlation in the clear condition and only a small to moderate decrease in correlation in the noise condition.

As we mentioned above in Subsection 8.1.2, we used the correlation results in selecting the specific forms of the weighting factor for each of the four weighting methods. Before we provide a list of the "best" objective measures, we present two results that helped us by reducing the number of objective measures we needed to monitor. First, Table 8 gives the 5-item correlation values for the four unweighted error measures: percent VUV error, percent UVV error, percent gross pitch error, and total error, which is the sum of the first three error measures. Of the first three error measures, the table shows that VUV error produced the highest correlation value for LPC/Clear and HDV/Clear. The same result was obtained in the study reported in [21], which considered only the LPC/Clear condition. For the two noise conditions, however, the VUV error produced not only the lowest correlation but also a positive correlation associating a higher error with a higher subjective rating, which is clearly wrong. This result and many others we came across in our investigation raise the caution that an objective measure that works well in the clear may not necessarily work well in the noise. Table 8 shows another important result that the total error always provided the highest correlation. Based on this result, we monitored only the total error in our subsequent work.

| Error | LPC/Clear | HDV/Clear | LPC/Noise | HDV/Noise |
|---|---|---|---|---|
| Percent VUV Error | -0.737 | -0.720 | 0.261 | 0.461 |
| Percent UVV Error | -0.503 | -0.540 | -0.509 | -0.662 |
| Percent Gross Pitch Error | -0.602 | -0.604 | -0.612 | -0.732 |
| Total Error | -0.957 | -0.964 | -0.813 | -0.761 |

TABLE 8. 5-item correlation results for four basic or unweighted error measures.

| Form of Energy Weighting | LPC/Clear | HDV/Clear | LPC/Noise | HDV/Noise |
|---|---|---|---|---|
| RMS Value | -0.984 | -0.990 | -0.770 | -0.602 |
| RMS Value in dB | -0.948 | -0.952 | -0.827 | -0.745 |
| RMS/MAX.RMS | -0.985 | -0.988 | -0.759 | -0.588 |

TABLE 9. 5-item correlation results for three forms of energy weighting.

Second, we compared the three forms of energy weighting: 1) RMS value, 2) RMS value in dB, and 3) RMS/MAX.RMS (see Subsection 8.1.2). Table 9 gives the 5-item correlation values for the four coder/background conditions. The first form produced the highest correlation in the clear, and the second form did in the noise. Consistent with our above-stated objective of achieving high correlation in the clear, we decided to use the RMS value for energy weighting in our subsequent work.

In our subsequent work, we monitored the correlation data for 125 objective measures of total error, which were the result of using no weighting or one or more of the four weighting methods with each of the three (VUV, UVV, and gross pitch) errors. We also monitored the correlation for the TI measure. Below, we shall use the simple notation, defined towards the end of Subsection 8.1.2, to refer to these objective measures.

From the correlation results obtained in the eight different conditions, we selected the 12 best objective measures that produced high correlation in the clear (-0.9 or better) and moderate-to-high correlation (-0.75 or better) in the noise. The average correlation over all eight conditions for each of

these 12 measures was high and ranged from -0.891 to -0.942. We present the 5-item correlation results in Table 10 and the 40-item correlation results in Table 11. We have ordered the 12 measures in terms of the average correlation over the eight conditions; the average correlations are given in Table 12. For comparison purposes, we have also given in Tables 10-12 the correlation results for the unweighted measure C-C-C and the TI measure. We see from Tables 10-12 that the TI measure performs quite well in the clear but quite poorly in the noise, producing an average correlation of only -0.561. Even the unweighted measure seems to be moderately robust under all eight conditions, with an average correlation of -0.824. The correlation results given in Tables 10-12 show our 12 best objective measures to be substantially more robust and yielding substantially higher average correlation as compared to the two reference measures. From Table 12, we see that as expected, the averages over the 5-item correlations were all larger than the averages over the 40-item correlations. If we consider only the overall ratings of the five pitch extractors, the average (5-item) correlation for the 12 best measures ranged from -0.906 to -0.982.

| No. | Objective Measure | LPC/Clear | HDV/Clear | LPC/Noise | HDV/Noise |
|---|---|---|---|---|---|
| **Best Measures:** | | | | | |
| 1 | EDFL-C-EF | -0.990 | -0.991 | -0.995 | -0.953 |
| 2 | EDFL-L-EDFL | -0.988 | -0.989 | -0.937 | -0.976 |
| 3 | EFL-DFL-EL | -0.986 | -0.995 | -0.958 | -0.937 |
| 4 | EDFL-EDL-EFL | -0.988 | -0.990 | -0.985 | -0.929 |
| 5 | EFL-DF-EL | -0.993 | -0.998 | -0.978 | -0.916 |
| 6 | EFL-EDFL-EL | -0.987 | -0.991 | -0.981 | -0.907 |
| 7 | EL-L-EDF | -0.994 | -0.996 | -0.961 | -0.864 |
| 8 | E-C-EDL | -0.995 | -0.999 | -0.902 | -0.826 |
| 9 | EFL-EFL-EL | -0.988 | -0.990 | -0.969 | -0.880 |
| 10 | E-EL-EDL | -0.984 | -0.984 | -0.903 | -0.805 |
| 11 | E-EDL-EF | -0.963 | -0.957 | -0.911 | -0.833 |
| 12 | E-EDL-EDL | -0.985 | -0.987 | -0.855 | -0.795 |
| **Reference Measures:** | | | | | |
| 13 | C-C-C | -0.961 | -0.968 | -0.817 | -0.770 |
| 14 | TI Measure | -0.992 | -0.989 | -0.185 | -0.015 |

TABLE 10. 5-item correlation results for 12 best measures and 2 reference measures.

| No. | Objective Measure | LPC/Clear | HDV/Clear | LPC/Noise | HDV/Noise |
|---|---|---|---|---|---|
| Best Measures: | | | | | |
| 1 | EDFL-C-EF | -0.929 | -0.924 | -0.867 | -0.888 |
| 2 | EDFL-L-EDFL | -0.919 | -0.895 | -0.854 | -0.899 |
| 3 | EFL-DFL-EL | -0.893 | -0.911 | -0.854 | -0.899 |
| 4 | EDFL-EDL-EFL | -0.909 | -0.917 | -0.839 | -0.870 |
| 5 | EFL-DF-EL | -0.905 | -0.922 | -0.842 | -0.862 |
| 6 | EFL-EDFL-EL | -0.901 | -0.922 | -0.823 | -0.849 |
| 7 | EL-L-EDF | -0.932 | -0.927 | -0.802 | -0.812 |
| 8 | E-C-EDL | -0.927 | -0.932 | -0.846 | -0.818 |
| 9 | EFL-EFL-EL | -0.909 | -0.923 | -0.761 | -0.775 |
| 10 | E-EL-EDL | -0.917 | -0.931 | -0.838 | -0.817 |
| 11 | E-EDL-EF | -0.918 | -0.922 | -0.814 | -0.823 |
| 12 | E-EDL-EDL | -0.909 | -0.928 | -0.831 | -0.835 |
| Reference Measures: | | | | | |
| 13 | C-C-C | -0.860 | -0.843 | -0.721 | -0.650 |
| 14 | TI Measure | -0.918 | -0.878 | -0.332 | -0.180 |

TABLE 11. 40-item correlation results for 12 best measures and 2 reference measures.

| No. | Objective Measure | Average over 5-item Correlations | Average over 40-item Correlations | Average over all eight Conditions |
|---|---|---|---|---|
| **Best Measures:** | | | | |
| 1 | EDFL-C-EF | -0.982 | -0.902 | -0.942 |
| 2 | EDFL-L-EDFL | -0.973 | -0.892 | -0.932 |
| 3 | EFL-DFL-EL | -0.969 | -0.889 | -0.929 |
| 4 | EDFL-EDL-EFL | -0.973 | -0.884 | -0.928 |
| 5 | EFL-DF-EL | -0.971 | -0.883 | -0.927 |
| 6 | EFL-EDFL-EL | -0.967 | -0.874 | -0.920 |
| 7 | EL-L-EDF | -0.954 | -0.868 | -0.911 |
| 8 | E-C-EDL | -0.931 | -0.881 | -0.906 |
| 9 | EFL-EFL-EL | -0.957 | -0.842 | -0.899 |
| 10 | E-EL-EDL | -0.919 | -0.876 | -0.897 |
| 11 | E-EDL-EF | -0.916 | -0.869 | -0.893 |
| 12 | E-EDL-EDL | -0.906 | -0.876 | -0.891 |
| **Reference Measures:** | | | | |
| 13 | C-C-C | -0.879 | -0.769 | -0.824 |
| 14 | TI Measure | -0.545 | -0.577 | -0.561 |

TABLE 12. Average correlation results for 12 best measures and 2 reference measures.

## 8.3 Recommendations

It is clear from the results given in the last section that the 12 best measures should be recommended. We address in this section how one may select a subset of these measures. We discuss below two ways of making this selection.

First, besides the requirement of high correlation, we must seek objective measures with meaningful combinations of the weighting methods. We believe that such measures are likely to continue to be valid for evaluation situations that are different from the ones used in our investigation (e.g., different pitch extractors, different but sufficiently large speech databases, different noise conditions, etc.) We believe that an objective measure should use the same F (pitch frequency) and L (context) weighting for <u>both</u> VUV and UVV errors. For example, the measure EDFL-C-EF does not satisfy this criterion, while the measure EFL-EDFL-EL does. Of the 12 best measures given in Tables 10-12, the measures 3, 6-9 satisfy the above criterion and may therefore be recommended.

Second, we consider the ability of the objective measure in rank ordering the test pitch extractors in a way that approximates the rank ordering

provided by the mean subjective rating. For this issue, let us use the overall mean rating. One can compute the Spearman's rank-order correlation and choose the measures that produce high correlation. But, we did not do this. For cases where the objective ordering differs from the subjective ordering, it is desirable to examine how close are the objective scores for the out-of-order pitch extractors; if the corresponding subjective scores are also close to each other, we may consider the objective measure still acceptable. In practice, unless the objective scores for two pitch extractors are different by more than some amount, we may not want to conclude that one is better than the other. With this rank-ordering criterion in mind, we examined in detail, for each of the 12 best measures, the ordering of the five pitch extractors by the objective scores and the objective scores themselves and compared them with the corresponding subjective ordering and mean ratings, for the four cases: LPC/Clear, LPC/Noise, HDV/Clear, and HDV/Noise. The measure EDFL-C-EF (No. 1 in Tables 10-12) that produced the highest average correlation also yielded the correct ordering for all but the HDV/Noise cases; this is because the objective measure gives only one score for both LPC and HDV coders, but the subjective ratings reversed the order of the fourth and fifth pitch extractors between the LPC/Noise and HDV/Noise cases (see Section 6.4). Of the 11 remaining measures, one (No. 3 in Tables 10-12) produced the right ordering for the two clear conditions, and none of the other 10 measures

produced the right ordering even for one coder/background condition. Many of the measures yielded the least error for JSRU in the clear and the most or the second most error for AMDFD in the noise, which is not in agreement with the subjective ordering (JSRU second best in the clear and AMDFD third best in the noise). However, we found four objective measures that satisfied the approximate ordering criterion we stated above; these four measures are Nos. 3, 4, 6, and 9 in Tables 10-12. For the best rank-ordering measure stated above and the four measures just identified, we have given in Tables 13 and 14 the objective scores for the five pitch extractors, respectively, in the clear and in the noise. For comparison purposes, we have also given in the tables the data from the unweighted measure (C-C-C) and the TI measure.

From Table 13 and Fig. 4 (see Section 6.4), we see that the first two measures and the TI measure correctly predicted the subjective ordering of the five pitch extractors in the clear condition. The other four measures reversed the ordering of the two best pitch extractors, but provided the right ordering otherwise. Table 14 shows that only the first measure performed quite well in predicting the subjective ordering in the noise condition. The next four measures correctly predicted JSRU and ILS as the best and second best algorithms, but failed in different ways at predicting the subjective ordering of the other three pitch extractors. The unweighted measure and the TI measure performed significantly worse.

| Objective Measure | AMDFD | Gold | H-S | ILS | JSRU |
|---|---|---|---|---|---|
| EDFL-C-EF | 5.31 | 23.25 | 22.01 | 7.10 | 6.01 |
| EFL-DFL-EL | 5.08 | 19.89 | 15.59 | 6.50 | 5.30 |
| EDFL-EDL-EFL | 4.95 | 19.61 | 18.97 | 6.91 | 4.48 |
| EFL-EDFL-EL | 4.67 | 15.96 | 15.22 | 6.38 | 3.93 |
| EFL-EFL-EL | 4.53 | 15.77 | 15.21 | 6.35 | 3.90 |
| C-C-C | 14.22 | 23.77 | 23.47 | 16.55 | 11.73 |
| TI Measure | 3.77 | 15.16 | 14.65 | 7.34 | 4.39 |

TABLE 13. Objective error scores produced by selected 5 best measures and 2 reference measures, for the clear condition.

| Objective Measure | AMDFD | Gold | H-S | ILS | JSRU |
|---|---|---|---|---|---|
| EDFL-C-EF | 28.54 | 32.10 | 30.01 | 28.02 | 16.21 |
| EFL-DFL-EL | 24.79 | 23.54 | 23.63 | 19.06 | 11.07 |
| EDFL-EDL-EFL | 30.58 | 31.83 | 29.57 | 28.27 | 16.15 |
| EFL-EDFL-EL | 21.31 | 23.06 | 19.99 | 18.71 | 11.04 |
| EFL-EFL-EL | 19.79 | 22.99 | 18.94 | 18.53 | 11.04 |
| C-C-C | 42.49 | 36.54 | 37.51 | 38.30 | 27.45 |
| TI Measure | 22.90 | 35.90 | 17.61 | 35.57 | 23.82 |

TABLE 14. Objective error scores produced by selected 5 best measures and 2 reference measures, for the ABCP noise condition.

## 9. SUMMARY AND FUTURE RESEARCH

From a review of the available pitch extractors, we chose and implemented five algorithms. By modifying an existing algorithm, we developed and tested an automatic method for extracting accurate, reference pitch and voicing data from the subglottal signal recorded simultaneously with the speech signal using a miniature accelerometer. Since the accelerometer is relatively insensitive to acoustic background noise, this method yields accurate pitch and voicing data even in noise.

For formal subjective evaluation of the chosen pitch extractors, we developed a speech database of 48 sentences that are likely to cause pitch and voicing errors, which facilitates efficient testing. We generated the test stimuli using two 2.4 kbit/s coders (LPC and HDV), 6 pitch extractors (5 algorithms under test and the reference), and 2 noise conditions (clear and ABCP). We ran two separate tests, one for each noise condition. Eight listeners rated the speech quality of the stimuli on an 8-point scale. The results of the subjective tests showed the reference subglottal-signal-based pitch extractor to be the best under all four coder/noise conditions, validating its use as reference in our subsequent objective evaluation work. We identified two best pitch extractors under test; one produced the highest

mean rating in the clear and the other, in ABCP noise.

The objective evaluation method we developed involves comparing, on a frame-by-frame basis, the test pitch extractor data with the reference pitch data, computing objective pitch and voicing error measures, and averaging over the sentences from the speech database. For developing objective measures, we first conducted a study to assess the perceptual effects of introducing different types and amounts of pitch and voicing errors into the reference pitch data. Based on the results of this study, we developed a large number of objective measures for evaluating pitch extractors, using different combinations of one or more of the following components: percentage of the frames containing voicing errors and gross pitch errors, energy weighting, weighting based on the duration of the errors, pitch frequency and pitch error weighting, and context-dependent error measurement. We also implemented two previously reported objective measures. We found that 12 of our objective measures provided consistently high correlation with mean subjective ratings in each of the four cases, two coders each in clear and in ABCP noise. In contrast, the previously reported measures provided high correlation in the clear and substantially lower correlation in the noise. Finally, our best overall objective measure produced excellent correlation, ranging from -0.953 to -0.995, with the overall mean subjective rating. This measure also

predicted nearly perfectly the rank ordering of the five test pitch extractors by the subjective rating, in all coder/noise conditions.

Finally, we suggest four problems for future work. First, the results from our detailed examination of the reference FPRDM pitch data given in Section 6.2 indicate the potential, through additional work, for substantially cutting down the 1% to 1.5% voicing errors. We believe that the FPRDM method can be and should be an excellent research tool in all speech processing work involving automatic extraction of accurate pitch or voicing or both.

Second, some of the pitch algorithms tested in this research (e.g., JSRU and FPRDM on speech; see Tables 3 and 4) can be improved with the use of a better voicing algorithm. Additional work may be performed by testing various ways of combining pitch extractors with voicing decision algorithms.

As a third area of work, we suggest testing the pitch algorithms in different acoustic backgrounds involving different noise spectra and different overall noise levels.

Fourth, the objective pitch evaluation measures developed in this research may be combined with other objective speech quality measures for

evaluating the speech quality of pitch-excited speech coders. Previous work in this area has not included pitch and voicing data as part of the objective speech quality measures, under the tacit assumption that these data have been extracted without any error (see [24], for example).

## REFERENCES


1. V. Viswanathan, M. Berouti, A. Higgins, and W. Russell, ``Development of a Good-Quality Speech Coder for Transmission Over Noisy Channels at 2.4 kb/s´´, Final Report No. 4916, Bolt Beranek and Newman Inc., March 1982, AD-A114068.

2. V. Viswanathan, M. Berouti, A. Higgins, and W. Russell, ``A Harmonic Deviations Vocoder for Improved Narrowband Speech Transmission´´, IEEE Int. Conf. Acoust., Speech, Signal Processing, Paris, France, May 1982, pp. 610-613.

3. T.E. Tremain, J.W. Fussell, R.A. Dean, B.M. Abzug, M.D. Cowing, and P.W. Boudra, Jr., ``Implementation of Two Real-Time Narrowband Speech Algorithms´´, Proc. EASCON ´78, Washington, DC, September 1978, pp. 698-708.

4. J. Wolf, K. Field, and W. Russell, ``Real-Time Implementation of the APC-SQ and LPC-10 Speech coding Algorithms´´, Final Report, No. 4855, Bolt Beranek and Newman Inc., June 1982, Contract No. DCA100-80-C-0034, AD-A116902.

5. V. Viswanathan, J. Makhoul, and A.W.F. Huggins, ``Speech Compression and Evaluation´´, Final Report No. 3794, Bolt Beranek and Newman Inc., April 1978, Contract No. MDA903-75-C-0180, AD-A055019.

6. W.J. Hess, Pitch Determination of Speech Signals: Algorithms and Devices, Springer-Verlag, Berlin, Heidelberg, New York, and Tokyo, 1983.

7. B. Gold, ``Computer Program for Pitch Extraction´´, J. Acoust. Soc. Amer., Vol. 34, No. 7, 1962, pp. 916-921.

8. B. Gold and L.R. Rabiner, ``Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain´´, J. Acoust. Soc. Amer., Vol. 46, No. 2, August 1969, pp. 442-448.

9. L.R. Rabiner, M.H. Cheng, A.E. Rosenberg, and A. McGonegal, ``A Comparative Study of Several Pitch Detection Algorithms´´, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 5, October 1976, pp. 399-413.

10. B.G. Secrest and G.D. Doddington, ``Postprocessing Techniques for Voice Pitch Trackers´´, IEEE Int. Conf. Acoust., Speech, Signal Processing, Paris, France, May 1982, pp. 172-175, Vol. 1.

11. H. Duifhuis, L.F. Willems, and R.J. Sluyter, ``Measurement of Pitch in Speech: An Implementation of Goldstein´s Theory of Pitch Perception´´, J. Acoust. Soc. Amer., Vol. 71, No. 6, June 1982, pp. 1568-1580.

12. J.L. Goldstein, ``An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones´´, J. Acoust. Soc. Amer., Vol. 54, No. 6, December 1973, pp. 1496-1516.

13. Signal Technology, Inc., ``ILS Application Note 1: Speech Analysis and Synthesis´´, 15 W. DeLaguerra Street, Santa Barbara, CA..

14. N. Green and B.C. Dupree, Joint Speech Research Unit , Personal Communication with Vishu Viswanathan.

15. B.G. Secrest and G.D. Doddington, ``An Integrated Pitch Tracking Algorithm for Speech Systems´´, IEEE Int. Conf. Acoust., Speech, Signal Processing, Boston, MA, April 1983, pp. 1352-1355, Paper No. 28.7.

16. W.H. Henke, ``Signals from External Accelerometers During Phonation: Attributes and Their Internal Physical Correlates´´, Quarterly Progress Report No. 114, Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, MA, July 1974, pp. 224-231.

17. K.N. Stevens, D.N. Kalikow, and T.R. Willemain, ``A Miniature Accelerometer for Detecting Glottal Waveforms and Nasalization´´, J. Speech and Hearing Research, Vol. 18, September 1975, pp. 594-599.

18. R.S. Nickerson, D.N. Kalikow, and K.N. Stevens, ``Computer-Aided Speech Training for the Deaf´´, J. Speech and Hearing Disorders, Vol. 41, February 1976, pp. 120-132.

19. V.R. Viswanathan, K.F. Karnofsky, K.N. Stevens, and M.N. Alakel, ``Multisensor Speech Input´´, Final Technical Report RADC-TR-83-274, Bolt Beranek and Newman Inc., December 1983, Contract No. F30602-82-C-0064.

20. V.R. Viswanathan, K.F. Karnofsky, K.N. Stevens, and M.N. Alakel,

``Multisensor Speech Input for Enhanced Noise Immunity to Acoustic Noise'', IEEE Int. Conf. Acoust., Speech, Signal Processing, San Diego, CA, March 1984, pp. 18A.3.1-18A.3.4, Vol. 2, Paper No. 18A.3

21. C.A. McGonegal, L.R. Rabiner, and A.E. Rosenberg, ``A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech'', IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 3, June 1977, pp. 221-229.

22. J.L. Flanagan, Speech Analysis Synthesis and Perception, Academic Press, New York, 1972.

23. D.H. Klatt, ``Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Pitch Perception'', J. Acoust. Soc. Amer., Vol. 53, No. 1, January 1973, pp. 8-16.

24. V.R. Viswanathan, W.H. Russell, and A.W.F. Huggins, ``Objective Speech Quality Evaluation of Real-Time Speech Coders'', Final Report No. 5504, Bolt Beranek and Newman Inc., February 1984, Contract No. DCA100-82-C-0005.

END

FILMED

11-84

DTIC